University of South Bohemia in České Budějovice

Faculty of Science

Bachelor's thesis

Nina Nenin

2019

University of South Bohemia in České Budějovice

Faculty of Science

NOVEL NON-CODING TRANSCRIPTS AT IMPRINTED LOCI IN MAMMALIAN OOCYTES AND EMBRYOS

Bachelor's thesis

Nina Nenin

Supervisor: Mgr. Lenka Gahurová, Ph.D.

Laboratory of early mammalian developmental biology, Department of molecular biology and genetics, Faculty of Science

České Budějovice

2019

Nenin N., 2019: Novel non-coding transcripts at imprinted loci in mammalian oocytes and embryos. Bc. Thesis, in English – 68 p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic

Annotation

Aims of this thesis were to annotate novel transcripts within clusters of imprinted genes in mouse oocytes and embryo, to analyze expression changes of these transcripts during mouse embryonic development, to identify enriched sequence motifs and potential transcription factors binding sites at promoters of transcripts within imprinted gene clusters as well as transposable elements acting as promoters of these transcripts and to identify potential candidates for further functional studies, using bioinformatic methods.

Affirmation

I hereby declare that I have worked on my bachelor's thesis independently and used only the sources listed in the bibliography. I hereby declare that, in accordance with Article 47b of Act No. 111/1998 in the valid wording, I agree with the publication of my bachelor thesis, in full to be kept in the Faculty of Science archive, in electronic form in publicly accessible part of the STAG database operated by the University of South Bohemia in České Budějovice accessible through its web pages. Further, I agree to the electronic publication of the comments of my supervisor and thesis opponents and the record of the proceedings and results of the thesis defense in accordance with aforementioned Act No. 111/1998. I also agree to the comparison of the text of my thesis with the Theses.cz thesis database operated by the National Registry of University Theses and a plagiarism detection system.

Date: Nina Nenin

Contents

1. Introduction

1.1. Genomic imprinting and imprinted genes

Genomic imprinting is an epigenetic process which affects a subset of genes in mammals and causes genes to be expressed in a monoallelic, parent-specific pattern. Monoallelic expression is controlled by epigenetic marks, predominantly DNA methylation, with differential occupancy on maternal and paternal allele. Primary differences in allelic DNA methylation are established in the germline (i.e. oocytes and sperm) as germline differentially methylation regions (gDMRs), and persist after fertilization during prenatal and postnatal development, leading to differential gene expression patterns from individual alleles. gDMRs which were functionally proven to regulate imprinting within the associated regions are referred to as imprinting control regions (ICRs) (Kelsey and Feil, 2013; Barlow & Bartolomei, 2014).

To this date, about 150 imprinted genes have been identified in mouse and mapped to 17 mouse chromosomes. It was shown that these genes tend to be clustered, specifically in 16 regions that contain two or more imprinted genes (Barlow & Bartolomei, 2014). The grouping of imprinted genes within clusters allows them to share common regulatory elements, such as gDMRs/ICRs and non-coding RNAs. Transcriptional regulation of the gene cluster by the differential methylation at ICR can be complex (illustrated in figure 1) : the methylated copy may repress one transcript and consequently promote the expression of other nearby genes, while the unmethylated ICR on the other allele acts as a promoter for a lncRNA and represses the expression of other genes in the cluster (Hanna and Kelsey, 2014; Andergassen et al. 2017). This is in contrast with direct regulation, when the unmethylated ICR promotes expression of the associated transcript.

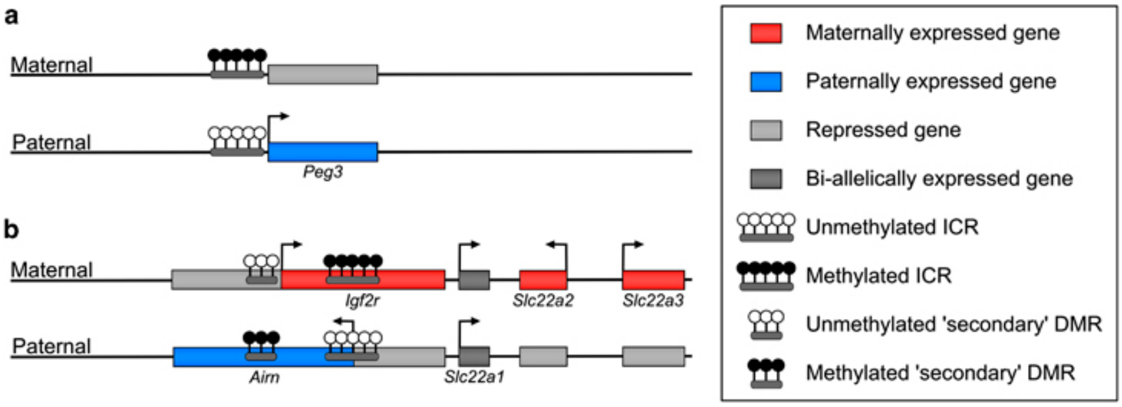

a
Maternal
Paternal
Peg3
b
Maternal
Igf2r
Slc22a2
Slc22a3
Paternal
Airn
Slc22a1
Maternally expressed gene
Paternally expressed gene
Repressed gene
Bi-allelically expressed gene
Unmethylated ICR
Methylated ICR
Unmethylated 'secondary' DMR
Methylated 'secondary' DMR

Figure 1. Examples of directly and indirectly regulated imprinted regions. Schematic representation of the (a) Peg3 imprinted gene on chromosome 7 and (b) the Igf2r imprinted cluster on chromosome 17. The illustration shows the expression status of genes on maternal and paternal alleles; horizontal arrows correspond to active promoters. (a) The differentially methylated ICR established during germ cell development is located at the promoter of the Peg3 gene and directly regulates the monoallelic transcription of this gene. (b) The maternally methylated ICR indirectly regulates the monoallelic expression of the adjoining genes at this locus, partially mediated by the monoallelic methylation acquired at the nearby secondary DMR at the Igf2r promoter. (Hanna and Kelsey, 2014;)

In addition to the effect on transcription, the other key characteristics of imprinting are the inheritance of imprints in somatic lineages through mitosis, initiation only in one of the two parental chromosomes, and the erasure in the early germ cells that must lose the inherited parental imprint in order for parental-specific identity to be established in the gametes (Ferguson-Smith, 2011).

Genomic imprints are classically defined as DNA methylation-dependent, yet the role of histone modifications and their relationship with DNA methylation and gene expression at imprinted loci is being studied. Although active and repressive histone modifications are known regulators of transcription, until recently, their role in imprinting was considered to be downstream of DNA methylation status of ICRs. (Ferguson-Smith, 2011). However, recent research suggests that histone modifications may play a role in the regulation of genomic imprinting. Particularly, trimethylation of histone 3 lysine 27 (H3K27me3) appears to regulate imprinted gene expression in extra-embryonic tissues independently of DNA methylation. This phenomenon is termed non-canonical imprinting (Hanna et al. 2019; Inoue et al. 2017).

It has been hypothesized that genomic imprinting evolved to play a certain role in mammalian development and reproduction, as imprinted genes are involved in various aspects of prenatal and postnatal development. In addition, due to imprinted genes affecting fetal growth, parthenogenesis in mammals is not possible. The evolution of imprinting is classically explained by the 'parental conflict' hypothesis. This hypothesis arises from the observations that embryonic growth is promoted by paternally expressed genes, while maternally imprinted genes repress fetal growth or minimize the effect of paternally expressed genes (Iwasa, 1998). Imprinted genes have also been identified in endosperm of some seed-baring plants, suggesting the importance of genomic imprinting in regulation of nutrient transfer; to this date, the reason behind this is still not known (Barlow & Bartolomei, 2014).

The importance of correct allele-specific expression of imprinted genes in mammalian development is exemplified by a number of human disorders that affect imprinted gene expression. In addition, the study of patients with such disorders, such as Beckwith-Wiedemann syndrome, Prader-Willi syndrome and Angelman syndrome, associated with parent-of-origin effects in their inheritance manner served as one of the key tools for the identification and understanding of the organization of imprinted genes. Studies conducted on patients and mouse as a model organism have been crucial for obtaining information about clusters of human genes, mapping ICRs and discovering epigenetic mechanisms that regulate genomic imprinting (Ferguson-Smith, 2011).

1.2.Establishment of genomic imprints in the germline

To this date, it has been shown that twenty gDMRs in imprinted regions have acquired methylation on maternal allele during oogenesis and only three during spermatogenesis. This indicates the differences in mechanisms by which DNA methylation marks are established in male and female gametogenesis. Imprints inherited from parents are erased in the embryonic germline by the combination of passive and active demethylation processes. During passive DNA demethylation, DNA methylation marks are gradually diluted as a consequence of repeated rounds of DNA replication without deposition of methylation marks on the newly synthesized DNA strand. Active DNA demethylation comprises the conversion of DNA methylation mark, 5-methylcytosine, to 5-hydroxymethylcytosine by the TET family of enzymes or deamination of 5-methylcytosine to thymine, where both 5-hydroxymethylcytosine and thymine can be removed and replaced by cytosine by base excision repair (Li and Zhang, 2014).

The re-establishment of DNA methylation (*de novo* DNA methylation) in unmethylated male and female germ cells occurs in different developmental stages and in different cellular contexts, and results in different DNA methylation patterns (Stewart et al. 2016). In the female gonad, DNA methylation is established after birth in meiotically arrested cells, while in the male gonad, *de novo* methylation occurs prior to meiosis in mitotically arrested prospermatogonia and the methylome has to be maintained during following mitotic proliferation and meiosis occurring between prospermatogonia and mature sperm. In sperm, almost all DNA is methylated with the exception of CpG-rich sequences (CpG is a DNA sequence where cytosine is followed by guanine) which are generally resistant do DNA methylation. In contrast, the methylome of oocytes is composed of large methylated and unmethylated domains, with methylated domains matching actively transcribed genes and unmethylated domains overlapping intergenic regions (Veselovska et al. 2015, Kobayashi et al. 2012).

The differential methylation of gDMRs between oocytes and sperm appears to be a consequence of the different overall methylation landscapes of the gametes. Maternally-methylated gDMRs colocalise with CpG-rich regions called CpG islands overlapping a promoter for coding on non-coding genes, whereas paternally-methylated gDMRs are located intergenically (Hanna et al. 2018).Although the maternally-methylated gDMRs overlap annotated promoters, transcription through these DMRs in the oocytes is a common feature of maternally-marked imprinted loci due to the activity of oocyte-specific upstream promoters (Chotalia et al. 2009, Veselovska et al. 2015).

1.3. Maintenance of genomic imprints after fertilization

After fertilization, the epigenetic landscape of both gametic genomes is reprogrammed with only a fraction of sequences keeping their methylation status from the gametes through the pre-implantation and later developmental stages. The paternal genome is rapidly demethylated through an active demethylation processes, while the maternal genome gradually loses most of its DNA methylation marks through passive demethylation during pre-implantation development. Exceptions from these global demethylation events are mostly imprinted gDMRs, which require different factors to prevent DNA methylation erasure. One such factor is maternal protein DPPA3 (also called PGC7/STELLA), which is highly expressed during oogenesis and persists in the pre-implantation embryo, and has a general role in protecting DNA from active demethylation by TET enzymes in early mouse embryo (up to the 2-cell stage) (Barlow & Bartolomei, 2014). Another factor that has more specific role in preserving gDMR methylation and is claimed to be an imprint specific factor is ZFP57 (Shi et al., 2019). Insights into the involvement of ZFP57 were obtained through gene targeting in mouse and from studies transient neonatal diabetes (TNDM). In patients with TNDM, loss of DNA methylation at multiple imprinted loci was detected, and it was associated with the mutation in the ZFP57 gene (Kelsey and Feil, 2013). In addition, results from the genetic experiments in mouse revealed that ZFP57 is essential for preventing the loss of DNA methylation at multiple imprinted loci. It has been shown that ZFP57 interacts with cofactor KAP1, which leads to the recruitment other repressive epigenetic regulators essential for maintenance of DNA methylation, for instance DNTMs, DNA binding factor UHRF/NP95 and H3K9 methyltransferase SETDB1 (Barlow & Bartolomei, 2014). One of the ZFP57-interacting proteins is the maintenance DNA methyltransferase DNMT1 that is consequently exclusively present at the imprinted loci and protects them from global passive DNA demethylation during pre-implantation development.

After fertilization, beside the loss of DNA methylation, histone modifications transmitted from the gametes are also reprogrammed in the pre-implantation embryo. Particularly, repressive H3K27me3 associated with non-canonical imprinting is lost during pre-implantation development and later re-established in the post-implantation embryo (Inoue, 2017; Hanna et al. 2019). The exact mechanism of this erasure and re-establishment of maternally-inherited H3K27me3 remain unclear (Hanna et al. 2019).

1.4.Imprinted gene clusters in mammals

As of now, 23 imprinted gDMRs and 96 imprinted genes have been identified, with an additional 13 putative imprinted genes in mouse placenta (Hanna et al. 2019). In the recent study conducted by Andergassen et al. (2017), it has been found that 19 out of 23 high confidence novel imprinted genes were in the vicinity of known imprinted genes, further indicating that the imprinted genes are regulated in clusters.

Imprinted regions have different sizes (up to 500 Mb or 500 kb) (Kaneko-Ishino and Ishino, 2019), containing both imprinted and non-imprinted genes, and both coding and non-coding transcripts. In addition, it appears that the size of the cluster regulated by the same ICR can differ between tissues - in placenta, ICRs regulate the imprinted expression of more distant genes than in classical somatic tissues. Therefore, it is not straightforward to estimate the borders of the clusters and to determine which genes are still controlled by the ICRs.

Gene expression is remodeled during development – it differs between oocyte, early embryos, late embryos, individual somatic tissues and placenta, involving genes in imprinted clusters as well. This is illustrated by the identification of novel oocyte-specific transcripts transcribed through the gDMRs in mouse oocytes that are not present in any embryonic stages or somatic tissues (Veselovska et al. 2015, Gahurova et al. 2017), regulating DNA methylation establishment at these loci. In addition, non-coding RNAs, such as *Kcnq1ot1*, are known important regulators in some imprinted loci (Mancini-Dinardo, 2006). Due to the tissue-specificity of non-coding RNAs, it cannot be excluded that some regulatory non-coding RNAs in imprinted regions were not identified yet.

Studies also show the difference in one locus during embryo development. For instance, it has been shown that there is lineage-specific regulation of Igf2r/Airn imprinted expression during gastrulation. In early embryonic development spreading of DNA methylation at Igf2r DMR2 during gastrulation was noted and at E6.5, both epiblast (Epi) and visceral endoderm (VE) lineages retain maternal ICR methylation. On the contrary, the epiblast expresses biallelic *Igf2r* and no *Airn*. However, both genes are imprinted in visceral endoderm of the same embryos indicating that there is a certain pathway distinctions that result in imprinted expression in VE but not in Epi at E6.5 – such as lineage-specific expression of chromatin binding/modifying genes established during preimplantation inner cell mass/trophectoderm differentiation. (Marcho et al. 2015).

Furthermore, many genes appear to be specifically imprinted in placenta in both human and mouse, suggesting that there are differences in transcriptional regulation between placenta and somatic tissues. Recent analyses revealed that some of the placenta-specific DMRs were associated with expression of imprinted genes such as *TIGAR, SLC4A7, PROSER2-AS1,* and *KLHDC10* (Hamada et al. 2016; Hanna et al. 2016). Recent studies identified novel imprinted transcripts in the vicinity of known imprinted genes predominantly specific for the placenta lineage (Andergassen et al. 2017; Hanna et al. 2019). One of such transcripts, within the *Slc38a4* locus, was shown to regulate the imprinted expression of *Slc38a4* (Hanna et al. 2019; Bogutz et al., under review).

Transposable elements (TEs) often act as promoters of oocyte-, embryo-, or placenta-specific transcripts (Veselovska et al. 2015; Franke et al. 2017; Macfarlan et al. 2012; Emera and Wagner, 2012). This also includes imprinted regions, where some TE-associated transcripts play important regulatory roles, such as providing transcription through gDMRs in the oocytes leading to their DNA methylation (Veselovska et al. 2015), or regulating the imprinting of the region in placenta lineage through yet unknown mechanisms (Hanna et al. 2019; Bogutz et al., under review). In non-imprinted genes, it was demonstrated that TEs can act as promoters for transcripts which act as enhancers and stimulate the transcription of nearby transcripts (Pi et al. 2010; Pi et al. 2017; Raviram et al. 2018). Therefore, it is possible that imprinted TE-associated transcripts regulate the imprinted expression of nearby genes in a similar manner, although it still remains to be elucidated what mechanism regulates the imprinted expression of TE-associated transcripts.

As of today, no study has globally described the transcriptional remodeling of gene expression in imprinted clusters during development. Considering that the transcriptomes of oocytes and embryos are not so well annotated as of somatic cells, due to the low amount of input material, it is possible that imprinted regions comprise some oocyte- or embryo-specific unannotated genes with potentially important roles. In this project, we therefore aimed to annotate all transcripts in imprinted regions in mouse, characterize their expression remodeling and shed more light on their transcriptional regulation.

2. Aims of the work

✓ Processing and mapping of publicly available RNA-seq datasets from various developmental stages and somatic tissues in mouse

✓ Annotation of transcripts within clusters of imprinted genes

✓ Analysis of expression changes of transcripts within imprinted gene clusters during development, and between embryonic and extraembryonic lineages

✓ Identification of enriched sequence motifs and potential transcription factors binding sites at promoters of transcripts within imprinted gene clusters

✓ Identification of transposable elements acting as promoters of transcripts within imprinted gene clusters

✓ Identification of potential candidates for further functional studies

3. Methods

The overall workflow of this project is visualized on fig. 2.

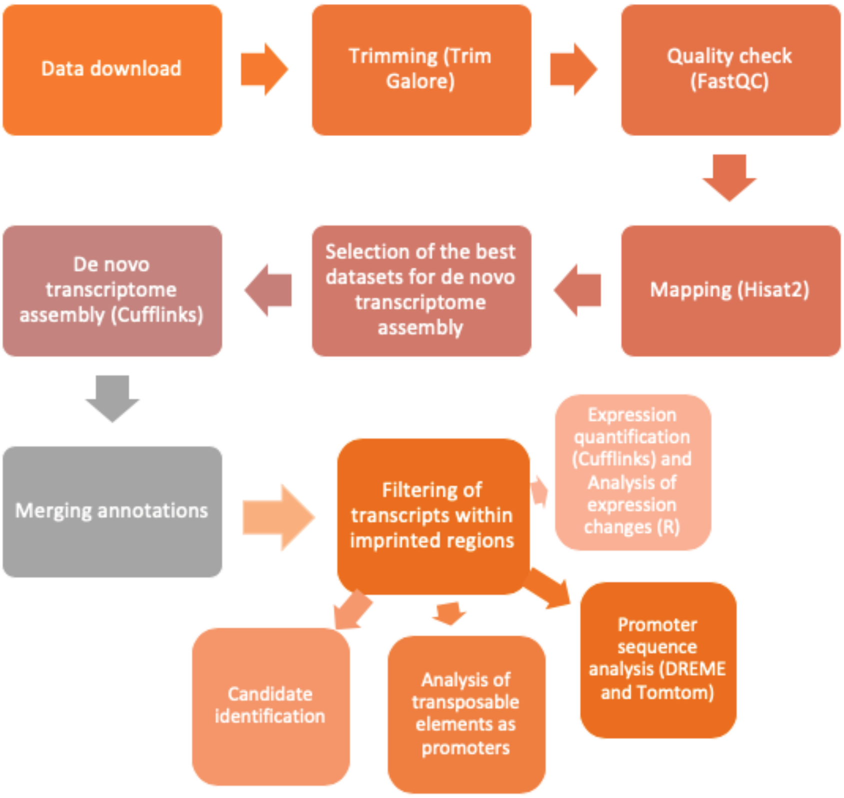

Data download
Trimming (Trim Galore)
Quality check (FastQC)
Mapping (Hisat2)
Selection of the best datasets for de novo transcriptome assembly
De novo transcriptome assembly (Cufflinks)
Merging annotations
Filtering of transcripts within imprinted regions
Expression quantification (Cufflinks) and Analysis of expression changes (R)
Promoter sequence analysis (DREME and Tomtom)
Analysis of transposable elements as promoters
Candidate identification

Fig. 2. Transcriptome analysis workflow

3.1.Datasets

RNA-seq datasets were searched for in NCBI Gene Expression Omnibus database and downloaded as fastq files from the European Nucleotide Archive (ENA, https://www.ebi.ac.uk). Datasets with following accession codes were used in this project: GSE70116, GSE71434, GSE98150, GSE76505, GSE75957, GSE124216. Detailed list of datasets used in this project can be found in Supplementary Table 1.

3.2. Trimming

To remove low-quality bases and adapters from the raw reads, program Trim Galore (www.bioinformatics.babraham.ac.uk/projects/trim_galore/) v0.4.1 was used with default parameters, specifying whether the reads were sequenced in single end or paired end mode. For single end reads, the command "trim_galore *fastq.gz" was used, for paired end reads, it was the command "trim_galore --paired *fastq.gz".

3.3. Quality control of trimmed reads

The quality of the trimmed reads (sequence quality and content, GC content, sequence length distribution, sequence duplication levels and overrepresented sequences) was checked using program FastQC (*http://www.bioinformatics.babraham.ac.uk/projects/fastqc*) v0.11.5 with default parameters, to check whether all the datasets are of sufficient quality for downstream analyses. The commands were "fastqc *_trimmed.fq.gz" and "fastqc *.fq.gz" for single end mode and for paired end mode, respectively.

3.4. Mapping

We mapped the trimmed reads to the previously indexed mouse GRCm38 genome (specified by -x parameter) using Hisat2 (Kim, et al. 2015; Pertea et al. 2016) v2.0.5 with parameters specifying the maximum and minimum values for soft-clipping per base (--sp) and modifying the output to be compatible with de novo transcriptome assembly using Cufflinks (--dta-cufflinks). The output file from Hisat2 with mapped reads (Sequence Alignment Map (sam) file), was further converted to Binary Alignment Map (bam) file using SAMtools view function of SAMtools v1.3.1 (H. Li, 2011; H. Li et al., 2009).

3.5. De novo transcriptome assembly and filtering

Prior to the de novo transcriptome assembly, the datasets that were split into two bam files (due to the sequencing in two different runs) were merged using SAMtools (http://samtools.sourceforge.net) v1.3.1 function merge (using the command samtools merge *merged.bam *rep1.bam *rep2.bam). All datasets were then sorted using SAMtools v1.3.1 function sort. Transcriptome assembly was done on sorted datasets using Cufflinks (http://cufflinks.cbcb.umd.edu/) v2.1.1. If multiple replicates were available for the same sample type, two or three replicates were selected for de novo transcriptome assembly based on read count, strand specificity and quality of the data (based on FastQC). After the assembly of the transcriptomes from the individual datasets, the annotations were merged into one final annotation using Cuffmerge function within Cufflinks v2.1.1 (- o option). This final annotation was further filtered using a Python v3.7 script previously developed in the laboratory (Supplementary file 1) to remove, based on genomic coordinates, transcripts not located within the imprinted regions. Furthermore, in program Seqmonk v1.44.0 (https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/), we compared the coordinates and strand specificity of known TEs with the annotated transcripts and we removed from the annotation all the transcripts overlapped by a same strand TE by more than 50%. This filtered annotation file (Supplementary file 2) was used in all downstream analyses.

3.6. Quantification expression of transcripts within imprinted clusters

To quantify the expression of transcripts we used Cufflinks v2.1.1 (command cufflinks -G mouse_merged_filtered.gtf -o output_folder sorted_mapped_reads.bam). The unit of expression is reads per kilobase of the transcript per million reads in the library (RPKM) for single read datasets and fragments per kilobase of the transcript per million reads in the library (FPKM) for paired end datasets.

3.7. Expression analysis, heat map and hierarchical clustering

First, we removed the transcripts with expression level under 0.1 in all datasets and mean-centered the values to be compatible with heatmap generation and clustering. This included logarithm transformation of values, the quantification of averages of these log2 values for each transcript from all the developmental stages and tissues and subtracting the average from each log2 value. These modifications were done in order to change the raw expression values into values reflecting the magnitude of expression changes between the datasets. From the list of transcripts and final mean-centered values we generated a .txt file for the hierarchical clustering and heatmap generation. Using RStudio (v.1.1.463) we did hierarchical clustering using function hclust and to visualize the expression profiles, we used heatmap. Using cutree command, we divided the transcripts into 20 main clusters (based on similarities of their expression profiles) and generated an output .txt file listing the number of the cluster for each transcript. List of all the functions used in RStudio is in Supplementary file 3. From the output file we quantified how many transcripts belong to each cluster and quantified average and median values for each of 11 clusters with more than 100 transcripts in all datasets using Microsoft Excel v16.29.1.

3.8. Sequence motif analysis

First, we obtained genomic coordinates (chromosome number, start and end base) of the promoters or wider regulatory regions of interest using program Seqmonk v1.44.0 (option Make probes, Upstream of feature, values 2000 bp upstream from the transcriptional start site, TSS, (+2000) and 500 bp downstream from TSS (-500), or +5000 and -5000 from TSS). When the probes were generated, we used option Fixed value quantitation (default settings) and saved the results as .txt file. From this file we selected only those transcripts which were used for the cluster analysis (with RPKM/FPKM value above 0.1 in at least one dataset), using Excel MATCH function. To obtain the sequences of the regions defined by coordinates, we used a Python script (v3.7) previously developed in the laboratory (Supplementary file 4) with mouse GRCm38 genome sequence. To find the enriched sequence motifs from the obtained sequences, we used program called DREME (Bailey, 2011) from MEME suite (Bailey et al. 2009, http://meme-suite.org/tools/dreme) using default settings. The analysis was performed for all regions, but also individually for each expression cluster with more than 100 transcripts. For each sequence (.txt file), we retained top 10 motifs (information about the letter code sequence, graphical logo, E-value, the numbers of positives and negatives). These motifs were then analyzed using Tomtom v5.0.5 (Gupta et al. 2007) program with default parameters to identify whether they match known binding sequence of a transcription factor, extracting the top 5 factors for each motif.

3.9. Transposable elements analysis

The analysis whether promoters of transcripts within imprinted regions are associated with TEs was done using Seqmonk (v1.44.0) and Microsoft Excel (v16.29.1). First, filtered .gtf annotation file was uploaded to Seqmonk together with files with the annotations of individual TE classes elements which were imported as reads. Import options were Column Delimiter Tab, Start at row 0, Chr Col 6, Start Col 7, End Col 8, Strand Col 10, the other options were left as default. Afterwards, we defined probes as +/-50 bp around TSS (Make Probes option, Upstream of feature was set to +50 and -50, the rest was left as default). When the probes were made, we quantified read counts within the probes (counting reads on the same strand as probe without any further modifications of the read counts). This allowed us to quantify whether TEs from each of TE classes overlap transcript promoter on the same DNA strand. The output report with read counts was saved as .txt file and imported in Excel (v16.29.1). Using MATCH function in Excel, we preserved the information only about promoters of transcripts which were used for the hierarchical clustering analysis. This was followed by the quantification of how many transcripts have value higher than 0 (Excel function COUNTIF) and therefore their promoters are associated with TE, for individual TE classes. This was analyzed for all transcripts, but also for individual expression clusters (using the Excel LOOKUP function and the information about clusters which was previously obtained).

3.10. Identification of potential candidates

For the identification of potentially interesting biological candidate transcripts, we used the quantification table (Supplementary table 2) to select transcripts which are relatively highly expressed in the oocytes and/or early embryos (above 0.5), and ideally with weak or no expression in somatic tissues. Transcripts should also be multiexonic and match the already annotated genes in the Ensembl annotation; to check that we used Seqmonk (v1.44.0). In addition to this, we also selected novel transcripts which do not match the already annotated genes in the Ensembl annotation. Lastly, we checked if there are multiple overlapping transcripts sharing the same exons, and if yes, whether the overlapping transcripts, or at least some of them, have similar expression profile as our selected transcript. The Seqmonk screenshots showing the expression of transcripts of interest were generated using Wiggle plot quantitation quantitating normalized read counts in 50 bp windows in Seqmonk (v1.44.0).

4. Results

4.1. Identification and processing of datasets

In order to generate a complete annotation of all transcripts within clusters of imprinted genes, including those that are novel and were not previously annotated, and to analyze expression changes of these transcripts across mouse development, we selected 41 publicly available RNA-seq datasets from oocytes, a range of embryonic stages, and neonatal and adult tissues (listed in Table 1) (Veselovska et al. 2015; Gahurova et al. 2017; Zhang et al. 2016; Wang et al. 2018; Zhang et al. 2018; Andergassen et al. 2017; Hanna et al. 2019). This comprises datasets from growing (postnatal day 5, 10 and 15 (d5, d10 and d15, respectively)) and fully grown (called germinal vesicle, GV) oocytes, all stages of early pre-implantation embryos, and individual embryonic cell lineages from the early blastocyst stage at embryonic day 3.5 (E3.5), late embryonic somatic tissues, neonatal brain and a number of adult somatic tissues from the major body organs. Embryonic cell lineages comprise the early segregating inner cell mass (ICM) and trophoectoderm (TE) and subsequent cell lineages of embryonic lineage (segregating from ICM), which gives rise to the embryo itself (epiblast - Epi, ectoderm - Ect, mesoderm - Mes, endoderm - End and primitive streak - PS), of lineage towards placenta (developing from TE, extraembryonic ectoderm - ExE), and an extraembryonic lineage segregating from ICM (visceral endoderm – VE) (Figure 3).

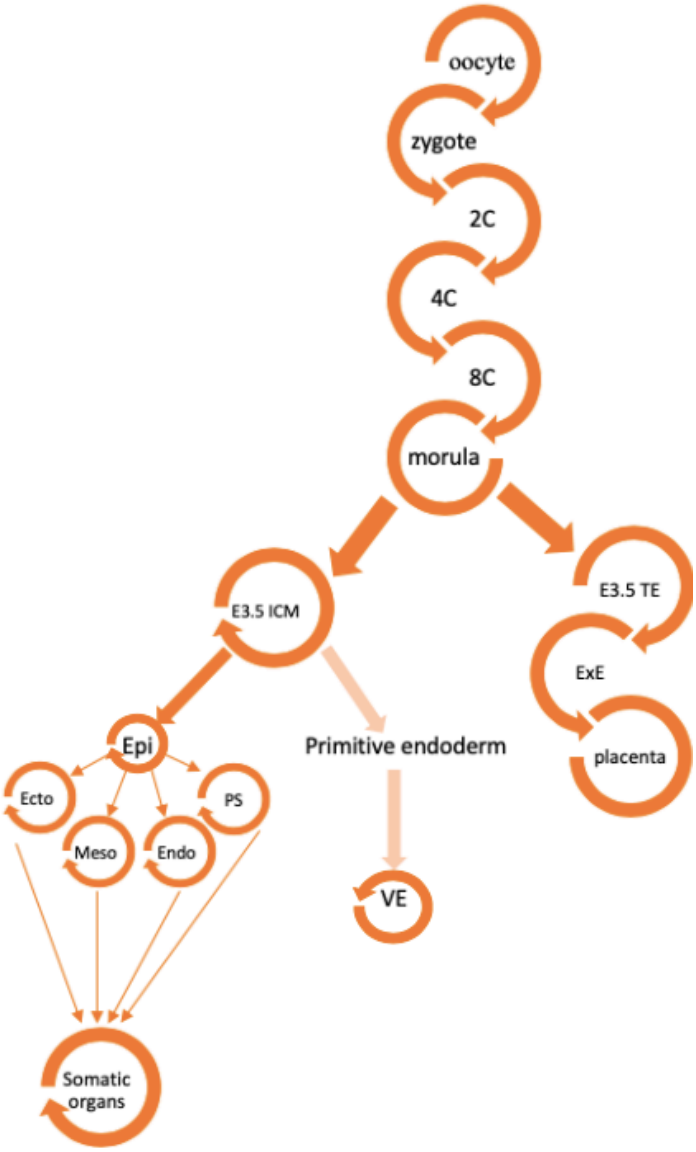


Figure 3. Embryonic cell lineages

After downloading the datasets, the adapters and bad quality bases were trimmed, and datasets were then quality checked and mapped to the GRCm38 mouse genome.

OOCYTE DATASETS	PRE-IMPLANTATION EMBRYONIC DATASETS	POST-IMPLANTATION EMBRYONIC DATASETS	SOMATIC TISSUES DATASETS
d5 oocytes	zygote	Ectoderm	D3 brain
d10 oocytes	late 2C embryo	Mesoderm	Adult brain
d15 oocytes	4C embryo	Endoderm	Adult liver
GV oocytes	8C embryo	E6.5 - Epi	Adult heart
	morula	E6.5 - ExE	Adult lung
	E3.5 - ICM	E6.5 - VE	Adult spleen
	E3.5 - TE	E7.5 - Epi	Adult thymus
	E4.0 - ICM	E7.5 -_ExE	Adult leg muscle
	E5.5 - Epi	E12.5 - placenta	Adult virgin mammary gland
	E5.5 – VE	E12.5_- liver	Lactating mammary gland
	ESC		Lactating brain
		E12.5 - VE	
		E16.5 - brain	
		E16.5 -_heart	
		E16.5 - liver	
		E16.5 -_placenta	
		Primitive streak	

Table 1. List of used datasets

4.2. Generating an annotation of transcripts within imprinted regions

After mapping the data, we performed de novo transcriptome assembly using Cufflinks (http://cufflinks.cbcb.umd.edu/) on the individual datasets. Then, the assembled annotations were merged using Cuffmerge within Cufflinks into one complete transcriptome annotation containing transcripts from all analyzed developmental stages. This final transcriptome annotation consists of 266340 transcripts.

From the final annotation, we were interested only in the imprinted regions. Based on the GeneImprint database (http://www.geneimprint.com) and recent publications (Andergassen et al. 2017; Inoue et al. 2017; Xu et al. 2011) we made a comprehensive list of all imprinted genes in mouse, containing 151 imprinted genes organized in 52 regions (some of these regions consist of only one imprinted gene). The genomic coordinates of imprinted clusters were defined by the first protein-coding gene with known function on either side that is either shown to be expressed bi-allelically, or with unknown imprinting status (Supplementary table 3). Using these coordinates of the borders of imprinted clusters, we filtered the assembled annotation to preserve only the transcripts inside these 52 imprinted regions, everything else was removed. The annotation file after filtering consists of 12307 transcripts.

4.3. Hierarchical clustering and expression analysis

To analyze the expression profiles of transcripts within the imprinted regions, the expression levels of transcripts were quantified using Cufflinks and the expression levels were averaged across replicates of the same dataset and across datasets from the same developmental stage from different sources. We removed transcripts overlapped by same strand transposable elements by more than 50%. We removed such transcripts because they are likely not to be real independent transcripts, just expressed transposable elements. This is in contrast with independent transcripts that use transposable elements as their promoters - but in these cases, the overlap with same strand transposable element is smaller than 50%, and they are often, but not always, spliced. In addition, we also removed all the transcripts with very low expression level (RPKM or FPKM under 0.1) in all datasets, as the expression changes, and expression itself, in such transcripts might be just due to the random transcriptional noise.

Then, we performed hierarchical clustering analysis which clusters transcripts with similar profile of expression changes across datasets and visualized the expression profiles using heatmap (Figure 4). In the heatmap, each row represents one transcript, columns represent datasets, high expression is visualized in yellow and low expression in blue.

After visual inspection of the heatmap, we decided to categorize the transcripts into twenty expression clusters. The heatmap also shows that the expression of transcripts in E7.5 Epi and Exe is predominantly higher than in other datasets, while the expression in End, Ect, Mes, and PS is relatively low. This might be related to the quality of the datasets and it probably does not represent the real biological situation. Therefore, we did not consider the expression patterns in these datasets as strongly as in the other datasets. We quantified how many transcripts belong to each cluster and quantified their average and median expression values. There are eleven clusters with more than 100 transcripts (Table 2), and we focused on them for further expression analysis.

The line graphs visualizing the average expression levels across datasets (figure 5a and 5b) show that transcripts are predominantly specific for a certain developmental stage. Cluster 3 represents oocyte-specific transcripts, which are degraded by 4C stage embryos (and never become highly expressed again), while in cluster 2 the expression peaks in the oocyte and in placenta, and in cluster 8 the transcripts are expressed in the oocytes and early embryo. In clusters 1 and 4, transcript expression is the highest in preimplantation embryos (in cluster 4, the expression also appears to peak in liver datasets). Clusters 6, 7 and 10 represent transcripts that are highly expressed in all postnatal somatic tissues and in late embryonic datasets with decreased expression in VE in all clusters and placenta in cluster 7. In cluster 5, there are transcripts expressed predominantly in placenta and other late embryonic datasets (from E9.5), but not at earlier embryonic stages or in postnatal tissues, whereas cluster 16 appears to represent brain-specific transcripts and cluster 9 contains transcripts with the highest expression in late embryonic tissues except brain and heart.

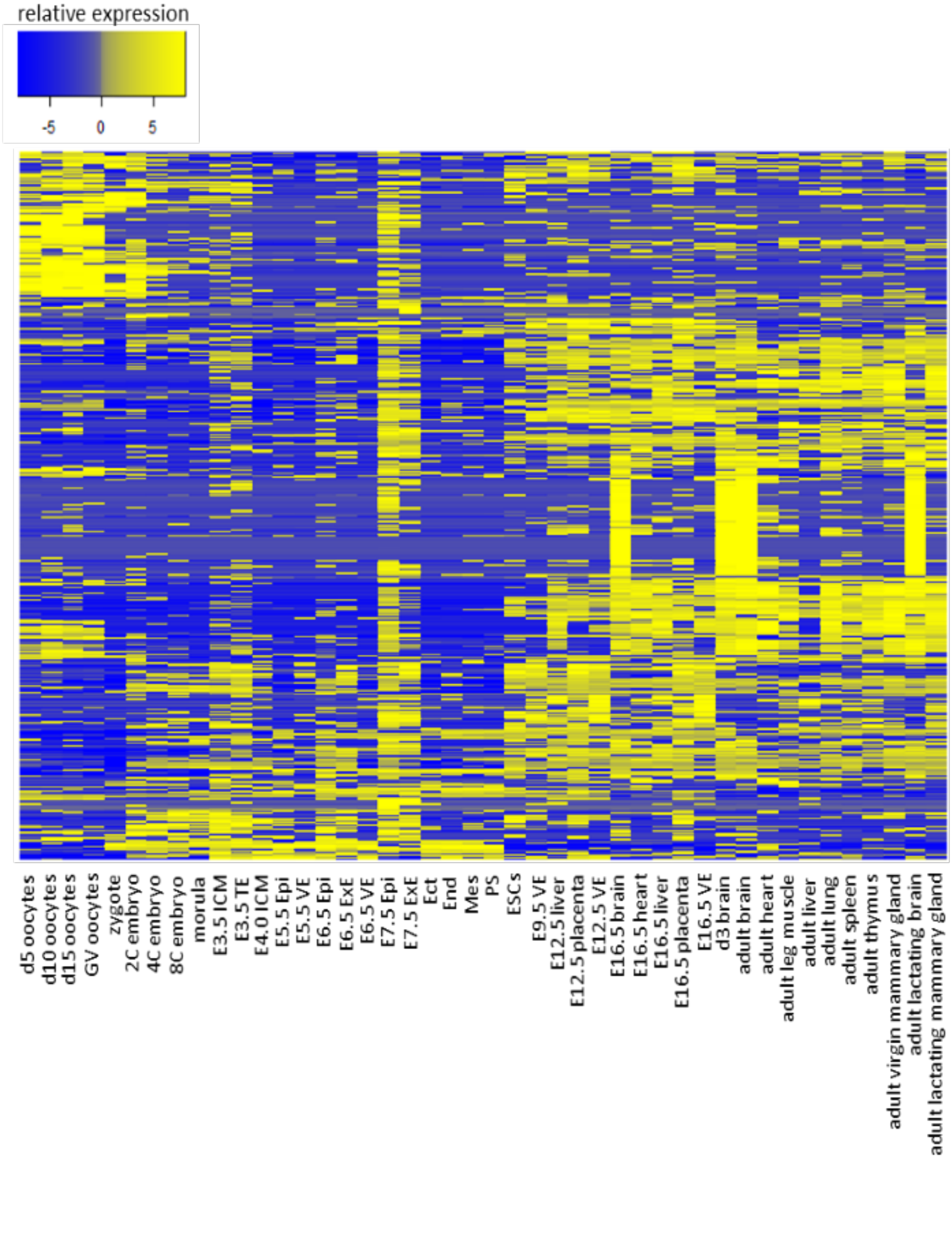


Figure 4. Heat map visualization of expression profiles. Each row represents one transcript and columns represent datasets; relative expression is visualized in shades of yellow (high expression) and blue (low expression)

Cluster	Number of transcripts
cluster_1	180
cluster_2	130
cluster_3	447
cluster_4	101
cluster_5	101
cluster_6	627
cluster_7	406
cluster_8	123
cluster_9	217
cluster_10	147
cluster_16	508

Table 2. Number of transcripts within each of the clusters that have 100 or more transcripts

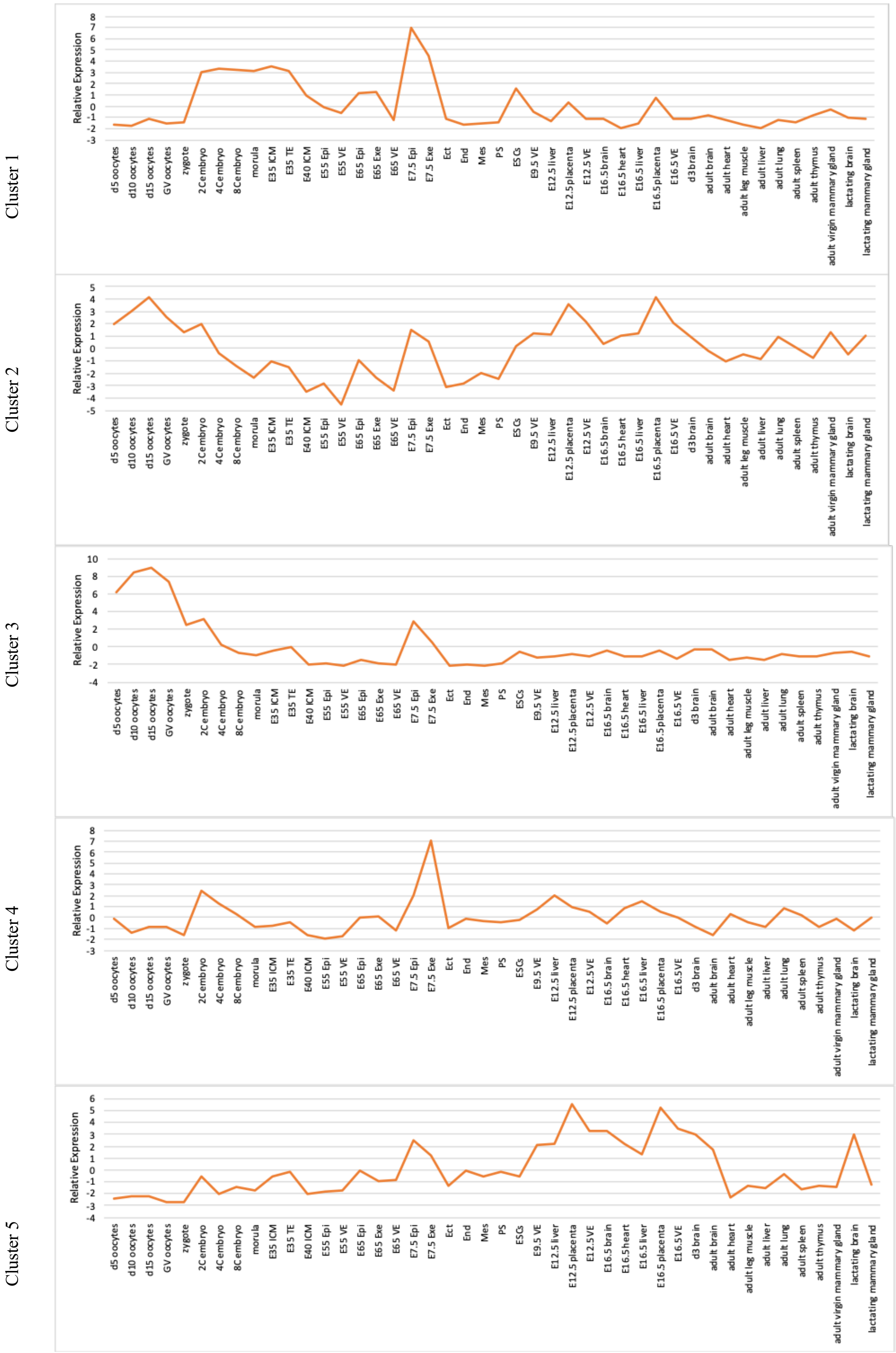


Figure 5a. Line graphs with average expression values (log2 transformed RPKM or FPKM)

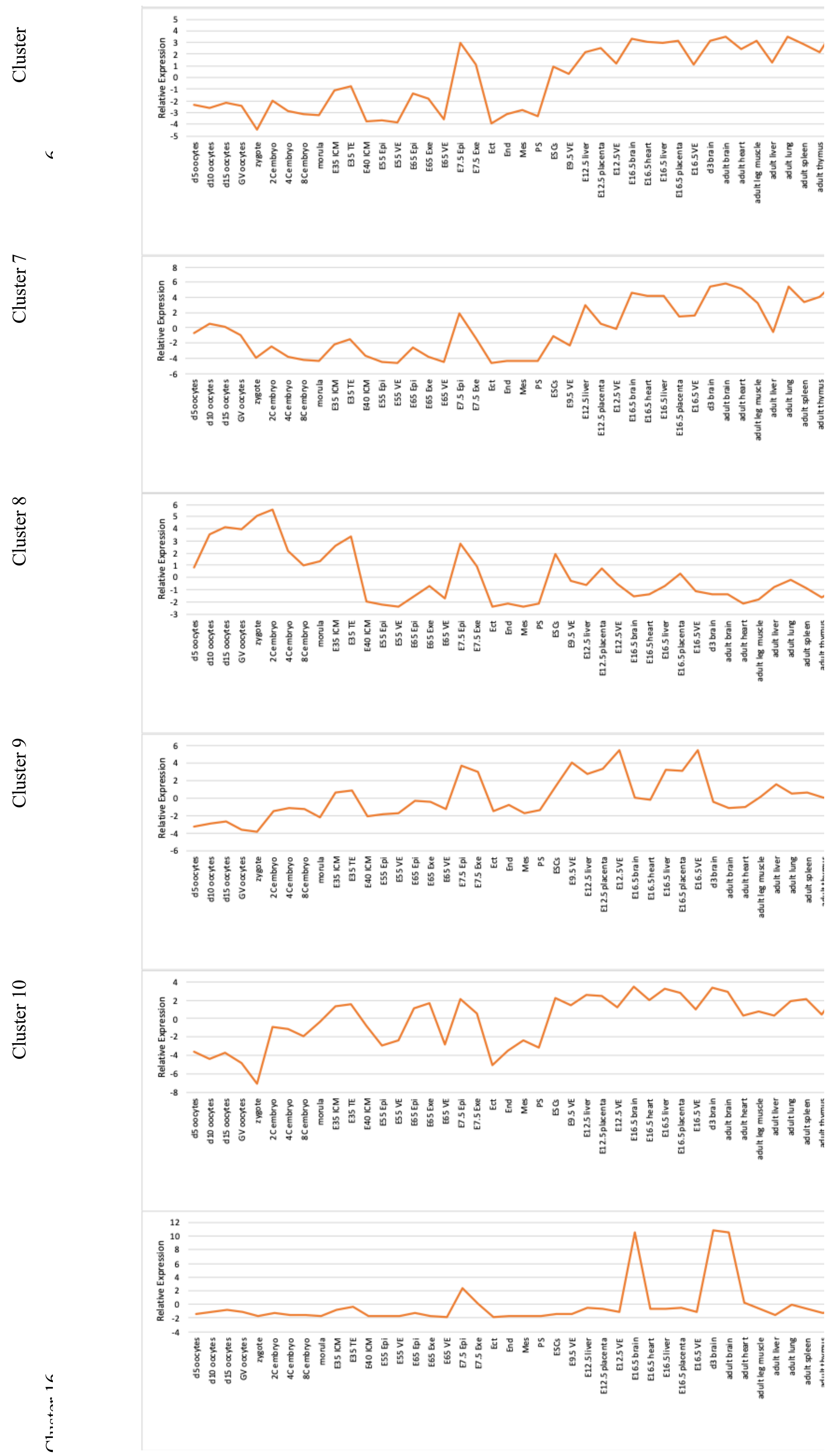


Figure 5b. Line graphs with average expression values (log2 transformed RPKM or FPKM)

4.4. Sequence motif analysis

In the analysis, we aimed to identify enriched sequence motifs that would indicate potential binding of transcription factors regulating expression of these transcripts. We looked for motifs in promoter sequences (-2000/+500 bp around the TSS), which are the primary binding sites of regulatory transcription factors, but also in broader regulatory regions (+-5000bp around the TSS), as transcription factors can also bind to a closely positioned enhancer regions.

To be able to see if there are different transcription factors regulating different clusters, we extracted the sequences of regions of interest of transcripts from the individual clusters (clusters 1-10 and 16) using a Python script previously written in the laboratory.

To find the enriched sequence motifs, we submitted the extracted sequences in the program DREME followed by another tool TOMTOM which associated the sequence motifs with known transcription factor binding sites. The results are summarized in the Supplementary table 4. Despite the lineage specificity of expression profiles of transcripts in the individual clusters, the results showed predominantly non-specific transcription factors. We observed that binding sequences of some transcription factors were identified in many of the clusters, such as ZSCAN4, FOXC1/FOXC2, POU-family factors, ELF3 and TCF3 factors as well as various ZFP factors. In the GeneCards database (https://www.genecards.org), these transcription factors are associated with the regulation of embryonic development, but also other developmental processes. The regulatory sequences of transcripts in the clusters with oocyte-, preimplantation embryo-, placenta- or brain-specific transcripts did not show enrichment for respective specific TFs, with the few exceptions. For example, in the cluster 16 containing transcripts enriched in brain datasets, we identified the enrichment for binding sites for TFAP2E (also called TCFAP2E), a transcription factor important for the development of central nervous system in humans (https://www.genecards.org/cgi-bin/carddisp.pl?gene=TFAP2E). However, the binding sites for this transcription factor were enriched also in the majority of other clusters (namely clusters 1-5, 7, 9 and 10). On the contrary, binding sites for HIC1 were identified only in clusters 1 and 4 containing transcripts with highest expression in the preimplantation development. Nevertheless, this transcription factor has no known association with the regulation of preimplantation development (https://www.genecards.org/cgi-bin/carddisp.pl?gene=HIC1). In addition, in the cluster 8 with transcripts highly expressed in the oocytes, the motif for PBX3 was identified (except the promoter region of transcripts in cluster 8, the motif for this factor was identified only in the broader regulatory regions of cluster 10). PBX3 is highly expressed in the ovaries (https://www.genecards.org/cgi-bin/carddisp.pl?gene=PBX3) suggesting it can act as an ovary or oocyte transcription factor.

4.5. Transposable element analysis

The goal was to find out what proportion of transcripts is using transposable elements as their promoters, and if this differs between expression clusters. We were interested particularly in ERVK elements, as recent research is showing that ERVK-starting transcripts can be involved in so-called non-canonical imprinting (Hanna et al. 2019).

The results showed that out of all 2816 transcripts, 497 use TE as their promoter, and it was found that the most common classes of TEs seem to be MaLR, ERVK and LINE-L1 (Table 5, Figure 6). The highest proportions of TEs acting as promoters were in oocyte-specific and oocyte-enriched clusters 3 (34.6% of all promoters) and 8 (29.0%), while the clusters 5, 6 and 9 with transcripts with the highest expression in late embryonic and/or postnatal tissues have the lowest proportions of transcripts using TEs as promoters (8.3%, 8.6% and 8.4%, respectively) (Figure 7, Table 7).

Based on the association of ERVK elements with non-canonical imprinting in placenta, we hypothesized that transcripts with promoters associated with these elements may be relatively common in cluster 5 and potentially also clusters 6 and 9, which all contain transcripts with high expression in placenta. However, this was not the case as only 1 and 3 transcripts initiate from ERVK promoter in clusters 5 and 6, respectively. In cluster 9, only 8 transcripts use ERVK as their promoter, but it represents 53% of all TE-initiated transcripts. The highest number of ERVK-initiated transcripts is in oocyte-specific cluster 3, where it represents 33% of TE-initiated transcripts.

	all	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7	cluster 8	cluster 9	cluster 10	cluster 16
LINE-L1	90	2	1	11	1	2	7	27	3	4	3	29
LINE-L2	9	0	0	1	0	0	4	3	0	0	1	0
LTR-ERV1	15	3	1	3	1	0	1	0	0	2	1	3
LTR-ERVK	107	10	2	42	1	1	3	2	7	8	1	6
LTR-ERVL	31	4	3	3	3	0	5	1	3	0	2	2
LTR-MaLR	152	7	11	66	4	1	14	10	14	0	4	6
SINE-B2	20	4	0	1	0	1	4	3	0	0	2	4
SINE-B4	21	1	2	1	1	1	5	1	2	1	0	5

Table 5. Numbers of transcripts with TE promoter

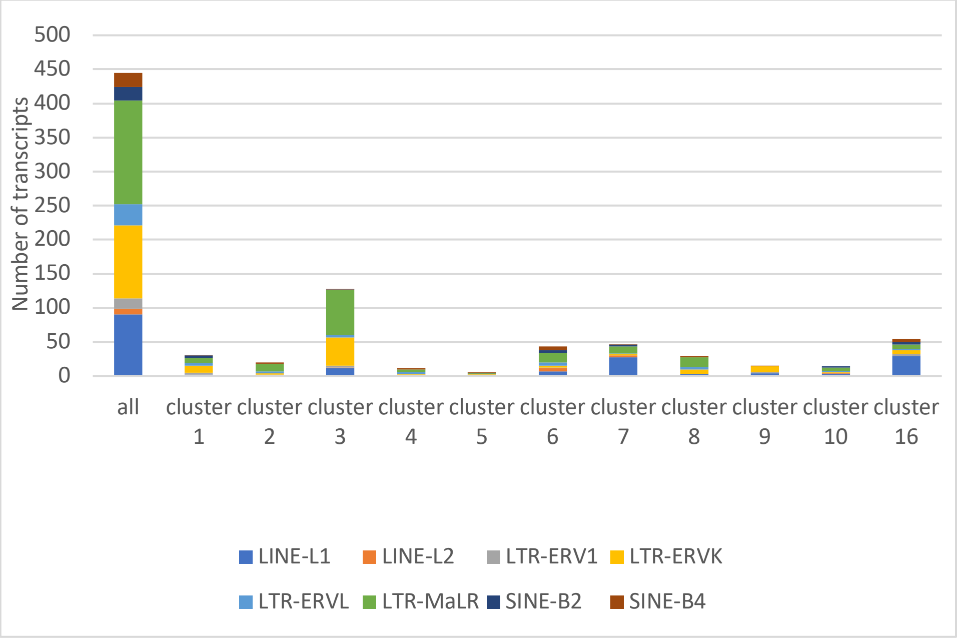


Figure 6. Numbers of transcripts with TE promoter

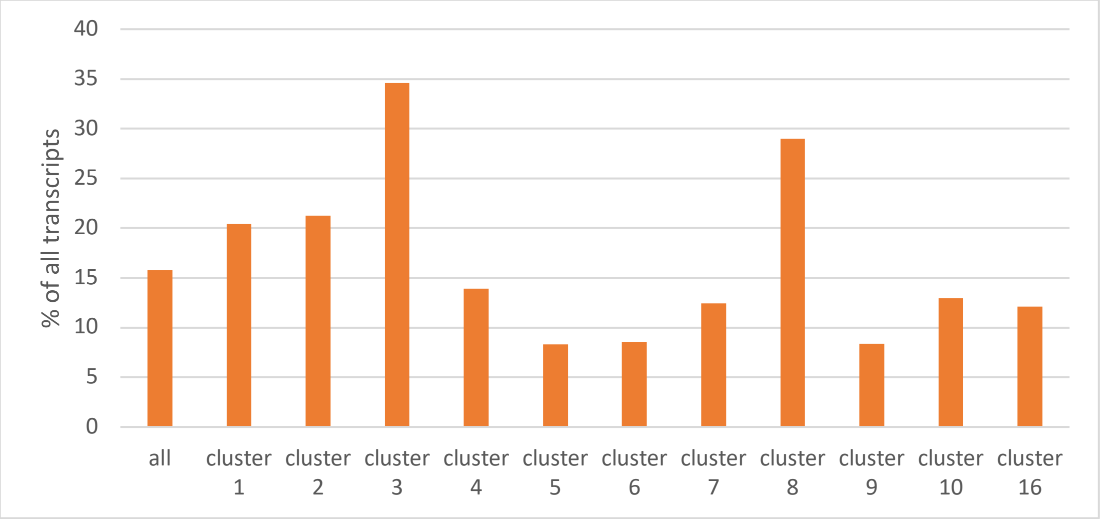


Figure 7. Proportions of transcripts using TE as their promoter

All	15.80256
cluster 1	20.39474
cluster 2	21.2766
cluster 3	34.59459
cluster 4	13.92405
cluster 5	8.333333
cluster 6	8.548708
cluster 7	12.43386
cluster 8	29
cluster 9	8.379888
cluster 10	12.96296
cluster 16	12.08791

Table 7. Proportions (%) of transcripts using TE as their promoter

4.6. Identification of potential candidates

In order to identify potential novel regulators of oocyte and/or embryonic development, we selected ten candidate transcripts that match the already annotated genes in the Ensembl annotation (Table 8) and ten novel candidates not overlapping annotated genes (Table 9). These transcripts can be further functionally tested in the laboratory by their downregulation and assessment of the phenotype in the oocytes or embryos.

The expression levels of candidate transcripts in the oocyte, preimplantation embryos up to early blastocyst stage and in somatic tissues are visualized in bargraphs (Figure 8 and Figure 9). We were particularly interested in the novel, previously not annotated transcripts. Based on expression profiles, novel candidates 1-6 are oocyte-specific, candidate 7 and 9 both belong to cluster number 1 which has transcript mostly expressed in preimplantation embryo, while candidates 8 and 10 are expressed in oocytes and early embryo (both belong to cluster 8). We visualized the expression of these transcripts in Seqmonk using wiggle plot quantification pipeline generating normalized read counts per 50bp windows, five examples (candidate transcripts 1, 2, 4, 7 and 8) are shown in figures 10-14, respectively.

	Transcript	Gene	Chromosome	Start	End
1.	TCONS_00090578	Galnt6	15	100690969	100729376
2.	TCONS_00100104	Arid1b	17	4993946	5348092
3.	TCONS_00100517	Tcp1	17	12916475	12922732
4.	TCONS_00105982	Map3k4	17	12227597	12316489
5.	TCONS_00106052	Wtap	17	12964461	12992622
6.	TCONS_00106119	Tcte2	17	13716427	13761386
7.	TCONS_00133034	H13	2	152669534	152704128
8.	TCONS_00192115	Mest	6	30738012	30752774
9.	TCONS_00224250	Osbpl5	7	143688023	143740341
10.	TCONS_00224262	Nadsyn1	7	143795489	143822841

Table 8. Candidate transcripts that are isoforms of known annotated genes and their genomic localization

	Novel transcripts	Chromosome	Start	End
1.	TCONS_00100057	17	3808903	4001766
2.	TCONS_00039018	11	22534962	22596030
3.	TCONS_00051879	12	109757023	109846872
4.	TCONS_00100409	17	9283117	9320122
5.	TCONS_00105849	17	7711656	7722073
6.	TCONS_00082993	15	72444718	72480405
7.	TCONS_00090036	15	97187271	97203784
8.	TCONS_00021986	10	96735954	96792734
9.	TCONS_00143954	2	168818237	168823232
10.	TCONS_00143956	2	168854718	168856989

Table 9. Candidate novel transcripts and their genomic localization

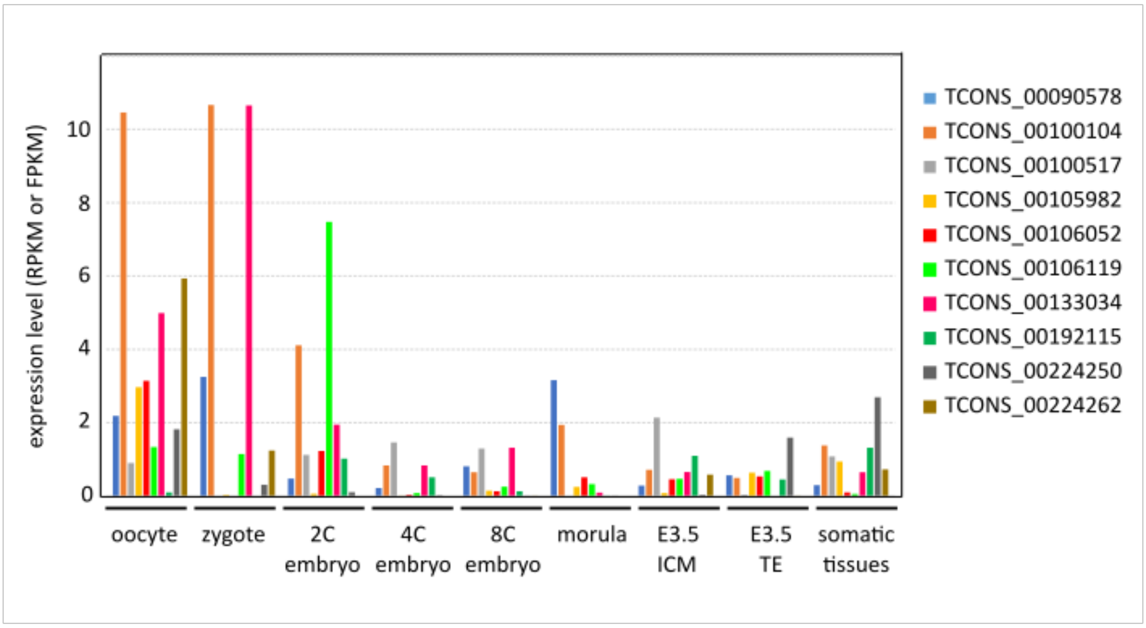


Figure 8. Expression levels of candidate transcripts that are isoforms of known genes

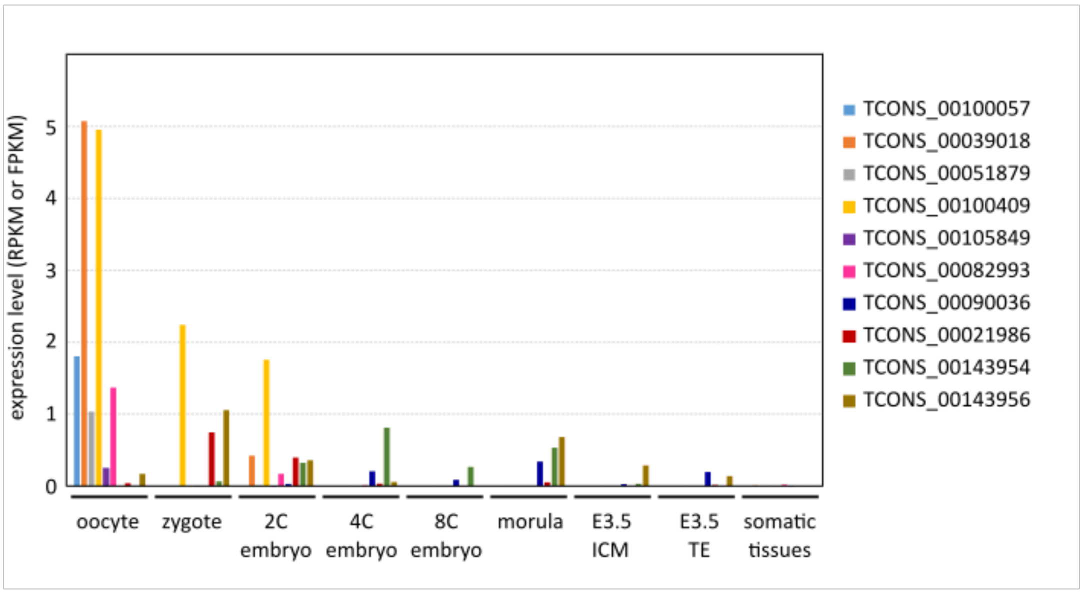


Figure 9. Expression levels of novel candidate transcripts

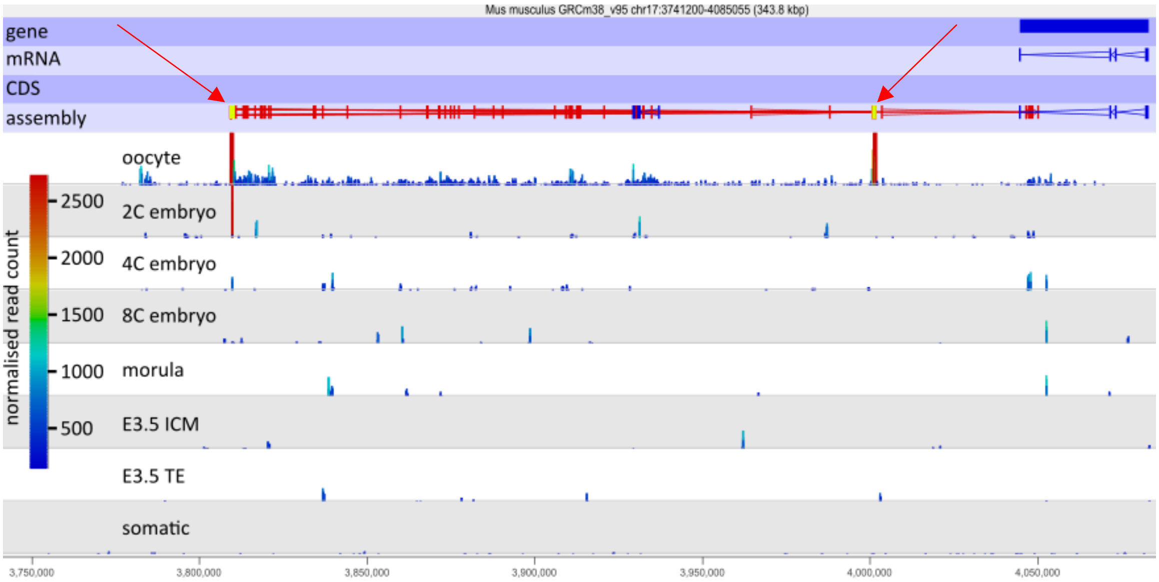


Figure 10. Visualized novel candidate transcript number 1 and its expression levels (its exons in de novo assembled transcriptome, marked as assembly, are highlighted in yellow and by red arrows, other exons belongs to other transcripts; rows marked as gene, mRNA and CDS show official annotation; expression levels are quantified as normalized read counts per 50 bp windows).

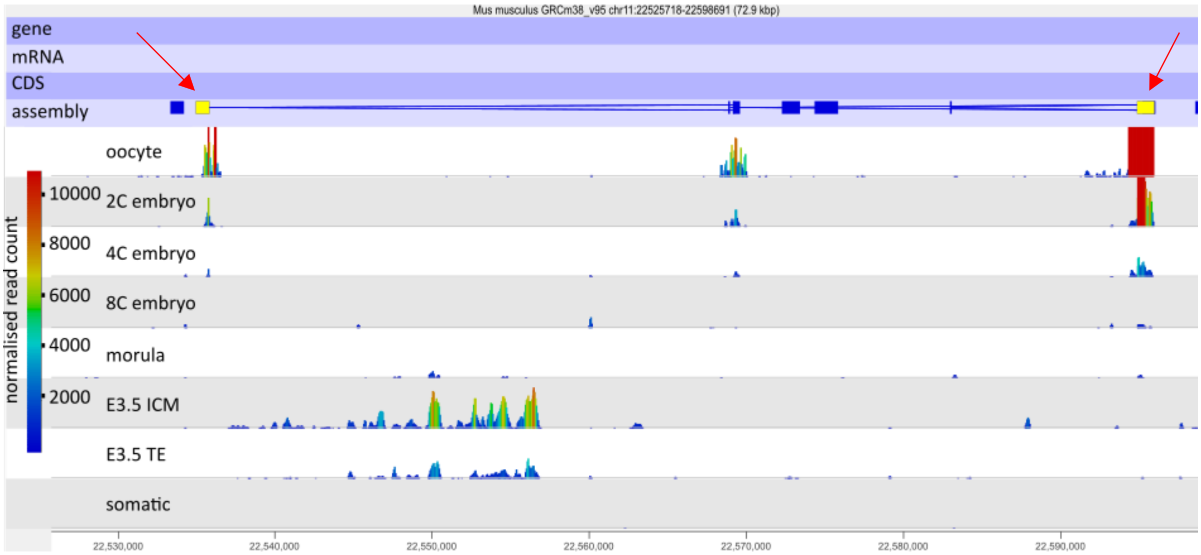


Figure 11. Visualized novel candidate transcript number 2 and its expression levels (its exons in de novo assembled transcriptome, marked as assembly, are highlighted in yellow and by red arrows, other exons belongs to other transcripts; rows marked as gene, mRNA and CDS show official annotation; expression levels are quantified as normalized read counts per 50 bp windows).

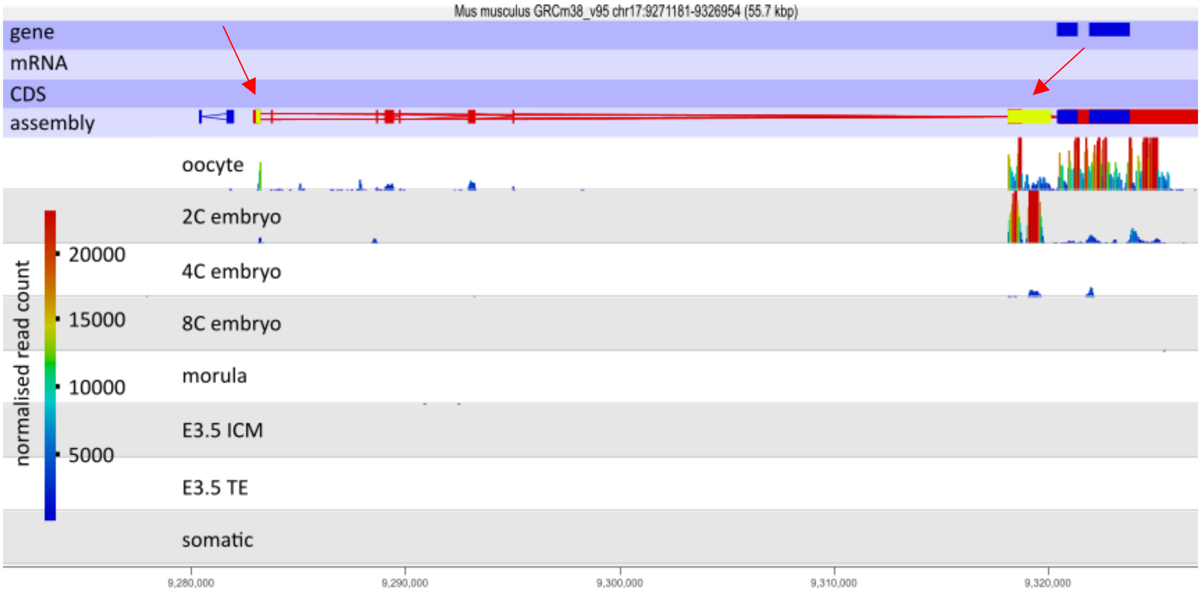


Figure 12. Visualized novel candidate transcript number 4 and its expression levels (its exons in de novo assembled transcriptome, marked as assembly, are highlighted in yellow and by red arrows, other exons belongs to other transcripts; rows marked as gene, mRNA and CDS show official annotation; expression levels are quantified as normalized read counts per 50 bp windows).

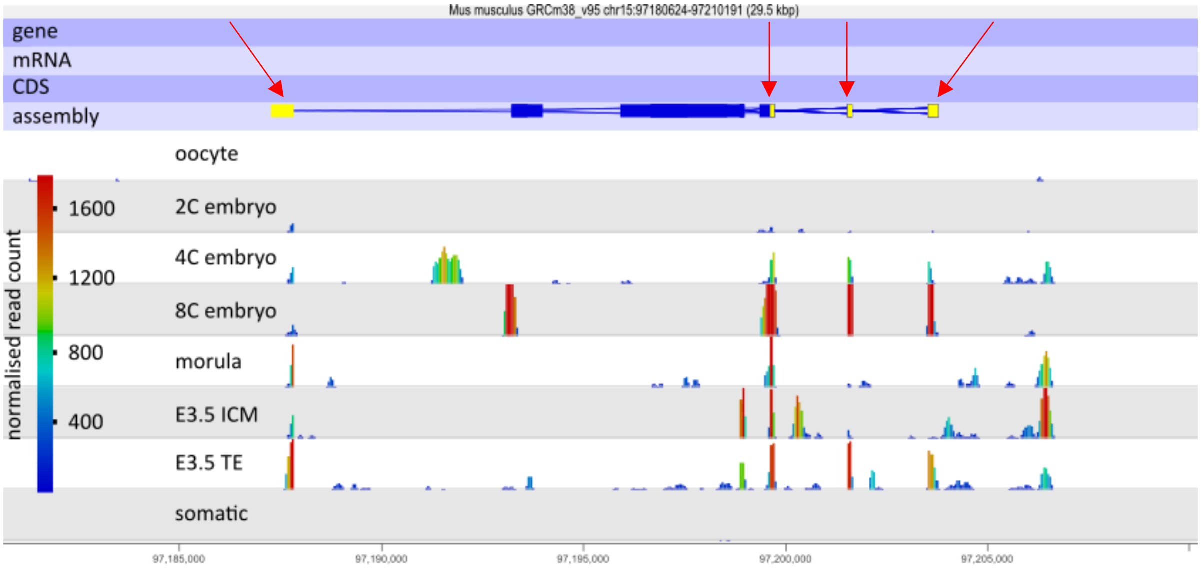


Figure 13. Visualized novel candidate transcript number 7 and its expression levels (its exons in de novo assembled transcriptome, marked as assembly, are highlighted in yellow and by red arrows, other exons belongs to other transcripts; rows marked as gene, mRNA and CDS show official annotation; expression levels are quantified as normalized read counts per 50 bp windows).

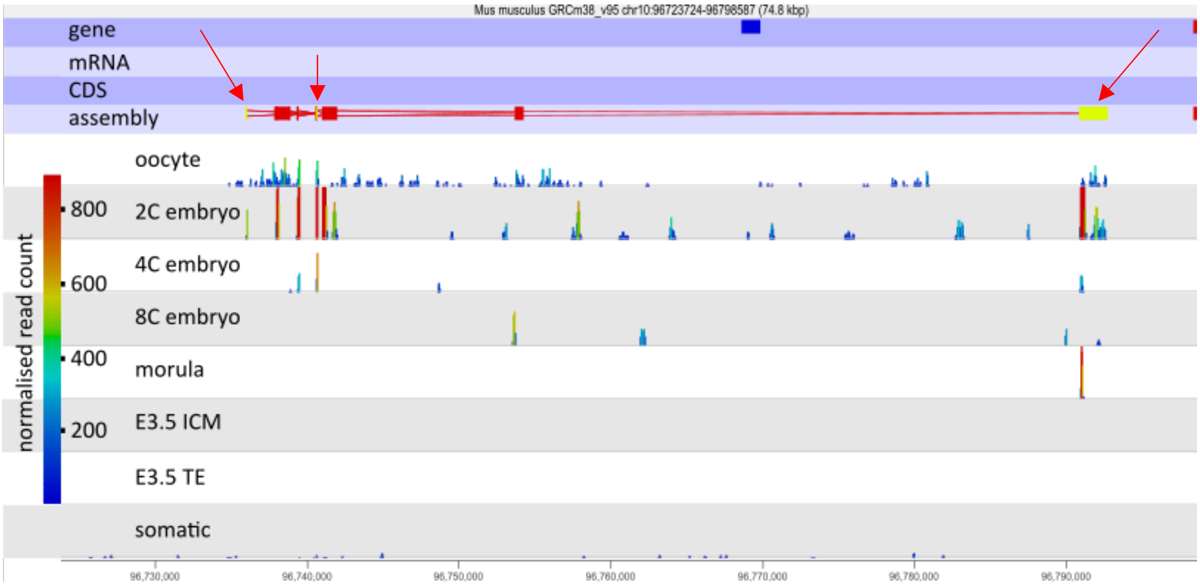


Figure 14. Visualized novel candidate transcript number 8 and its expression levels (its exons in de novo assembled transcriptome, marked as assembly, are highlighted in yellow and by red arrows, other exons belongs to other transcripts; rows marked as gene, mRNA and CDS show official annotation; expression levels are quantified as normalized read counts per 50 bp windows).

5. Discussion

In this thesis, we downloaded, processed and mapped 41 publicly available RNA-seq datasets from mouse oocytes, embryos and somatic tissues. Using these data, we assembled the complete transcriptome of imprinted regions across mouse development. We found out that transcripts within imprinted regions are often specific for certain developmental stage (or stages) and a substantial number of them appears to be specific for oocytes or embryonic development. We further noticed that there do not appear to be specific transcription factors regulating the expression of transcripts with similar expression profiles, but that a substantial proportion of transcripts, particularly those that are oocyte-specific, employs TEs as their promoters. We selected 20 transcripts expressed specifically in the oocytes or early preimplantation embryos for further functional analysis in the laboratory.

To date, this is the first assembly of the complete transcriptome of mouse imprinted regions, annotating a substantial number of novel, previously not annotated transcripts. The fact that these transcripts were not previously annotated might be due to their tissue- or lineage-specificity. This agrees with other studies performing de novo transcriptome assembly from low input samples – for example, previous de novo assembly of whole oocyte transcriptome also identified a high number of novel transcripts (Veselovska et al. 2015, Gahurova et al. 2017). The novel transcripts are likely to be long non-coding RNAs, as these are often tissue- or cell type-specific (Zhu et al. 2016). From adult somatic tissues, the only tissue that differed from the others was brain, with a considerable number of brain-specific transcripts (this agrees with Andergassen et al. (2017) that brain has specific imprinting differing from other main body organs). Some transcripts are highly expressed in placenta, agreeing with the fact the placenta has high number of placenta-specific imprinted genes (Hanna et al. 2016, Inoue et al. 2017, Andergassen et al. 2017).

To this date there is no optimal strategy for generating RNA-seq libraries, and approaches also depend on the amount of starting material which is generally scarce for oocyte and embryos, but abundant for somatic tissues. Therefore, the quality of individual datasets used in this thesis differed and it is probably reflected in the quality of transcriptome assembly of respective datasets. For example, transcriptome assembly from the oocytes (Veselovska et al. 2015 datasets) and late embryonic and postnatal tissues (Andergassen et al. 2017 datasets) is probably more precise (better reflecting the reality) than from early embryos, as Veselovska et al. (2015) and Andergassen et al. (2017) datasets are deeply sequenced and are strand specific, while early embryonic datasets mostly lack the strand information.

Without the strand specificity of the reads, the direction of the de novo assembled transcript (if it is encoded on plus or minus DNA strand) cannot be correctly estimated.

The number of annotated transcripts is probably higher than the real number of transcripts, as some monoexonic genes are likely to be part of nearby multi- or mono-exonic genes, just that read density was not high enough to connect them (as in Veselovska et al., 2015). Also, some genes are likely to have fewer isoforms than annotated, as some isoforms can be just artefacts of the assembly.

Moreover, the imprinted regions defined in this thesis might not be accurate, knowing that imprinted regions controlled by the same imprinted gDMR have different sizes based on the tissue/cell type (Andergassen et al. 2017). In order to circumvent this, we tried to include the broadest regions; also, as borders, we just took first gene that is either known to be not imprinted, or with unknown imprinted status. Therefore, there is small likelihood that these genes might be imprinted in some less studied tissues.

The main limitation of the expression profiling analysis was that the datasets were generated by different sources, therefore, they can vary a lot due to the technical reasons (mostly the due to the differences in approaches and kits used for RNA extraction, cDNA synthesis and library preparation, and the number and length of reads, causing different sequencing depth). In the heatmap, we can see that some datasets were generally showing lower expression levels for all genes (Ect, End and Mes), and some generally high (E7.5 Epi, E7.5 Exe) - that is probably due to the mentioned technical differences in library preparation and sequencing depth.

Sequence motif analysis did not identify anything of particular interest - even for genes expressed predominantly in the oocytes, placenta or brain, we generally did not find oocyte-, placenta- or brain-specific transcription factor binding sites. This might be either due to the real lack of specific regulator, or due to their binding further than 5kb from the annotated TSS, or due to the imprecise annotation of TSSs and promoters.

TE analysis revealed that TEs are used as promoters mostly in transcripts highly expressed in the oocytes. Interestingly, their proportion is similar to the proportion of TE-associated transcripts as identified in overall oocyte transcriptome in Veselovska et al. (2015), suggesting that imprinted regions are not particularly enriched in or depleted for at least oocyte-specific TE-associated transcripts. In addition, our results results agree with Veselovska et al. (2015) that TEs acting as promoters are mostly LTR-MaLR elements. Furthermore, we did not identify a high number of ERVK-initiated transcripts, associated with placenta-specific non-canonical imprinting (Hanna et al. 2019, and Bogutz et al., under review), not even in clusters with transcripts highly expressed in placenta.

This thesis serves as a basis for future research investigating the level of species-specificity of imprinted transcripts and other transcripts in imprinted regions. The existing manuscript (Bogutz et al., under review) describes the species specificity of some imprinted transcripts expressed in the oocytes of mouse, rat and human. By performing de novo transcriptome assembly and expression analysis in other mammalian species we can expand the study to more species and especially more developmental stages. The species specificity is also hypothesized by the use of TEs as promoters - those transcripts with TE promoters are more likely to be specific only for species where those TEs are present. In addition, if some transcripts are conserved between species, we can analyze whether their expression profile is the same in different species. By selecting only ERVK-initiating transcripts, we can explore non-canonical imprinting, its conservation between species and its apparent specificity for placenta lineage. Moreover, the analysis presented in this thesis will be improved by differentiating between known and novel transcripts, and between imprinted, unknown and bi-allelically expressed transcripts (particularly in terms of their expression profiles and TE analysis). As a different future direction, we selected candidate transcripts with interesting expression profiles, which could be tested for their functions (by their downegulation, knock-out, over-expression etc.) in oocyte and pre-implantation embryos.

6. Conclusion

In this project, we for the first time assembled the complete transcriptome of mouse imprinted regions across development using publicly available RNA-seq datasets. This led to the identification of a number of novel, previously unannotated, transcripts, with potential functional roles in development or in the regulation of imprinting in the respective imprinted gene cluster. Transcripts in the imprinted regions appear to be mostly expressed in a specific developmental stage or period. Their expression does not differ largely between adult somatic tissues with the exception of brain expressing a considerable number of brain-specific transcripts. Despite the cell type- or developmental-specificity of the transcripts, they do not appear to be regulated by specific transcription factors. A substantial proportion of transcripts highly expressed in the oocytes or preimplantation embryos uses transposable elements as promoters, particularly LTR-MaLR elements, however, the frequency does not differ from the overall proportion of transposable elements-initiated transcripts in the oocytes. Despite the association of ERVK-initiated transcripts with placenta-specific non-canonical imprinting, we did not identify a large number of placenta-specific ERVK-initiated transcripts. This suggests that there are probably not many more non-canonically imprinted transcripts than those few already identified. This project will serve as a basis for future research studying species-specificity of transcripts in imprinted regions across mammalian species, and their association with transposable elements. In addition, selected transcripts will be functionally tested in the laboratory for their potential functions in the oocytes and in embryonic development.

7. References

Aaron B. Bogutz, Julie Brind'Amour, Hisato Kobayashi, Kristoffer N. Jensen, Kazuhiko Nakabayashi, Hiroo Imai, Matthew C. Lorincz, Louis Lefebvre (2019) Evolution of imprinting via lineage-specific insertion of retroviral promoters. bioRxiv 723254; doi: https://doi.org/10.1101/723254

Andergassen D, Dotter CP, Wenzel D, Sigl V, Bammer PC, Muckenhuber M, Mayer D, Kulinski TM, Theussl HC, Penninger JM et al. Mapping the mouse Allelome reveals tissue-specific regulation of allelic expression. Elife 2017;6:e25125.

Barlow DP, Bartolomei MS. (2014) Genomic imprinting in mammals. Cold Spring Harb Perspect Biol 6: a018382. doi:10.1101/cshperspect.a018382

Chotalia, M., Smallwood, S. A., Ruf, N., Dawson, C., Lucifero, D., Frontera, M., James, K., Dean, W. and Kelsey, G. (2009). Transcription is required for establishment of germline methylation marks at imprinted genes. Genes Dev 23: 105-117.

Deena Emera, Günter P. Wagner, Transposable element recruitments in the mammalian placenta: impacts and mechanisms, *Briefings in Functional Genomics*, Volume 11, Issue 4, July 2012, Pages 267–276, https://doi.org/10.1093/bfgp/els013

Ferguson-Smith AC. (2011) Genomic imprinting: the emergence of an epigenetic paradigm. Nat. Rev. Genet. 12, 565–575.doi:10.1038/nrg3032

Franke V, et al. (2017). Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes. Genome research 27, 1384–1394

Gahurova, L. et al. (2017) Transcription and chromatin determinants of de novo DNA methylation timing in oocytes. Epigenetics Chromatin 10, 25.

Hamada H, Okae H, Toh H, Chiba H, Hiura H, Shirane K, et al. (2016). Allele-specificmethylome and transcriptome analysis reveals widespread imprinting in thehuman placenta. Am J Hum Genet.99:1045–58. https://doi.org/10.1016/j.ajhg.2016.08.021.

Hanna CW, Palacios RP, Gahurova L, Schubert M, Krueger F, Biggins L, Andrews S, Colome-Tatche M, Bourc'his D, Dean W, Kelsey G. (2019). Non-canonical imprinting in extra-embryonic tissues is driven by endogenous retroviral insertions. Genome Biology (2019) 20:225 https://doi.org/10.1186/s13059-019-1833-x

Hanna, C. W. et al. (2016) Pervasive polymorphic imprinted methylation in the human placenta. Genome Res. 26, 756–767. doi:10.1101/gr.196139.115

Hanna, C. W., and Kelsey, G. (2014). The specification of imprints in mammals. Heredity 113, 176–183. doi: 10.1038/hdy.2014.54

Hanna, C. W., Demond, H., & Kelsey, G. (2018). Epigenetic regulation in development: Is the mouse a good model for the human? Human Reproduction Update, 24(5), 556– 576. https://doi.org/10.1093/humupd/dmy021

Hu T., Pi W., Zhu X., Yu M., Ha H., Shi H, Choi J., Tuan D. (2017) Longnon- coding RNAs transcribed by ERV-9 LTR retrotransposon act in cis to modulate long-range LTR enhancer function. Nuc Acids Res 2017; E pub ahead of print; http://dx.doi.org/10.1093/nar/gkx055

Inoue, A. et al. (2017) Maternal H3K27me3 controls DNA methylation-independent imprinting. Nature 547, 419–424

Iwasa, Y. (1998). The conflict theory of genomic imprinting: how much can be explained Curr. Top. Dev. Biol.40, 255–293

Kaneko-Ishino Tomoko, Ishino Fumitoshi (2019) Evolution of viviparity in mammals: what genomic imprinting tells us about mammalian placental evolution. Reproduction, Fertility and Development 31, 1219-1227. https://doi.org/10.1071/RD18127

Kelsey,G. and Feil,R. (2013) New insights into establishment and maintenance of DNA methylation imprints in mammals. Philos. Trans. R Soc. Lond. B Biol. Sci., 368, 20110336.

Kobayashi H, Sakurai T, Imai M, Takahashi N, Fukuda A, Yayoi O, et al. Contribution of intragenic DNA methylation in mouse gametic DNA methylomes to establish oocyte-specific heritable marks. PLoS Genet. 2012;8:e1002440.

Macfarlan T.S. et al. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature, 487, pp. 57-63

Mancini-Dinardo, D., Steele, S. J., Levorse, J. M., Ingram, R. S. and Tilghman, S. M. (2006). Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. Genes Dev. 20, 1268-1282.doi:10.1101/gad.1416906

Marcho C, Bevilacqua A, Tremblay KD, Mager J. (2015). Tissue-specific regulation of Igf2r/Airn imprinting during gastrulation.Epigenetics Chromatin. 2015; 8:10. doi: 10.1186/s13072-015-0003-y.

Pi W, Zhu X, Wu M, Wang Y, Fulzele S, Eroglu A, Ling J, Tuan D. (2010). Long-range function of an intergenic retrotransposon. Proc NatlAcad Sci U S A. 107:12992–12997

Raviram, R.; Rocha, P.P.; Luo, V.M.; Swanzey, E.; Miraldi, E.R.; Chuong, E.B.; Feschotte, C.; Bonneau, R.; Skok, J.A.9 (2018). Analysis of 3D genomic interactions identifies candidate host genes that transposable elements potentially regulate. Genome Biol. 19, 216.

Shi et al. Epigenetics & Chromatin (2019) 12:49 doi:10.1186/s13072-019-0295-4

Shobhit Gupta, JA Stamatoyannopolous, Timothy Bailey and William Stafford Noble, "Quantifying similarity between motifs", *Genome Biology*, 8(2):R24, 2007

Stewart, K. R., Veselovska, L., & Kelsey, G. (2016). Establishment and functions of DNA methylation in the germline. *Epigenomics*, *8*(10), 1399–1413. doi:10.2217/epi-2016-0056

Timothy L. Bailey, "DREME: Motif discovery in transcription factor ChIP-seq data", *Bioinformatics*, 27(12):1653-1659, 2011.

Timothy L. Bailey, Mikael Bodén, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, William S. Noble, "MEME SUITE: tools for motif discovery and searching", *Nucleic Acids Research*, 37:W202-W208, 2009.

Veselovska L, Smallwood SA, Saadeh H, Stewart KR, Krueger F, Maupetit-Méhouas S, Arnaud P, Tomizawa S, Andrews S, Kelsey G. (2015) Deep sequencing and de novo assembly of the mouse oocyte transcriptome define the contribution of transcription to the DNA methylation landscape. Genome Biol 16: 209.

Zhu, J., Chen, G., Zhu, S. *et al.* Identification of Tissue-Specific Protein-Coding and Noncoding Transcripts across 14 Human Tissues Using RNA-seq. *Sci Rep* 6**,** 28400 (2016) doi:10.1038/srep28400

8. Supplementary files and tables

- Supplementary file 1. Python script used for gtf filtering
- Supplementary file 2. Filtered gtf file (on CD)
- Supplementary file 3. R script for the heatmap and clustering analysis
- Supplementary file 4. Python script used to extract sequences
- Supplementary table 1. List of datasets used in this project
- Supplementary table 2. Quantification table (on CD)
- Supplementary table 3. List of imprinted regions in the format chromosome:start-end
- Supplementary table 4. Summary of sequence motifs analysis results

Supplementary file 1. Python script used for gtf filtering (Written by Sylvia Ramirez, edited by Nikolas Tolar)

```
# Python code filters regions based on chromosome and specific start and end of
bases.
#and, makes an extra filtering for removing transcripts with one exon, from a
subset of genes.


import re
import os

input_filename = "rat_merged.gtf"

chom_start_end_file = "instructions.txt" #no header, 3 columns -
choromosome,start,end (separated by tabs)



'''

nik edit begins ------------------------
'''

chromosomes = []
bases = []

feed_file = open(chom_start_end_file,'r')

line = feed_file.readline()

while line != '':
    line_split = line.split('\t')
    chromosomes.append(line_split[0])
    bases.append([int(line_split[1]),int(line_split[2])])
    line = feed_file.readline()

'''
nik edit ends   ------------------------

'''

# creates output file name: input_filename + filtered.gtf
output_filename = input_filename[:input_filename.rfind(".")] +
"_exons_filtering_subset.gtf"

# opens the input file
with open(input_filename) as f:
    # reads all lines
    lines = f.readlines()

# closes input file
f.close()
# gets number of lines (used for progress)
count_lines = len(lines)

# initializes counter to 0 (used for progress)
counter = 0

# counter findings
findings = 0

# opens output file
#of = open(output_filename, "w")
```

```
#open temp file
of = open("temp.gtf", "w")

startAt_history = {}

def indexFrom(input_data, search_for, startAt):
    for i in range(startAt, len(input_data)):
        if input_data[i] == search_for:
            return i

def geneids_in_region():
    print("Initializing...")
    global counter, findings, startAt_history

    # if one transcript is within the region => set it to true
    for l in lines:

        counter += 1
        # splits line by tab and creates an array
        l_data = re.split(r'\t+', str(l))

        if l_data[2] == "exon":

            # checks if same chromosome (string)
            if l_data[0] in chromosomes:
                startAt_history[l_data[0]] = 0

                for x in chromosomes:
                    if x == l_data[0]:

                        startAt = 0
                        if l_data[0] in startAt_history:
                            startAt = startAt_history[l_data[0]]

                        index = indexFrom(chromosomes, l_data[0], startAt)

                        startAt_history[l_data[0]] = index + 1

                        b = bases[index]
                        l_start_base = b[0]
                        l_end_base = b[1]
                        # checks start position
                        if l_start_base <= int(l_data[3]) <= l_end_base:
                            # checks end position
                            if l_start_base <= int(l_data[4]) <= l_end_base:
                                of.write(str(l))
                                findings += 1

        # prints progress
        #print(input_filename + ": " + str(counter) + "/" + str(count_lines) + "
Found: " + str(findings))
        #if counter%10000 == 0:
            #print("First cleavage " + str(counter) +" Found: " + str(findings))


geneids_in_region()


# closes the output file
of.close()

##############################
#Additional filtering
#reading temporary file

with open("temp.gtf") as f:
    # reads all lines
    lines_tmp2 = f.readlines()

#remove temporary file
os.remove("temp.gtf")

#######
#reading exons that need to be remove
#######

import pandas as pd


dataset=pd.read_csv("rat_to_be_removed_if_1_exon.txt",delimiter="\t")

#create list with values
remove_marker = dataset["Probe"].tolist()

#make format as we have in our transcript_id list
for k in range(len(remove_marker)):
    remove_marker[k] = '"{0}"'.format(remove_marker[k])


print("\n\nThird cleavage start...\n")
transcript_id_2 = []

for i in range(len(lines_tmp2)):
    data_from_line = []
    #split line for extracting ids
    for j in lines_tmp2[i].split(";")[1:-1]:
        data_from_line.append(j.split(" ")[2])

    transcript_id_2.append(data_from_line[0])


# find delete from lists needed lines
count = 0
for l in range(len(remove_marker)):
#    if counter%1 == 0:
#        print("Found for removing: " + str(count))
    if transcript_id_2.count(remove_marker[l]) == 1:
         index = transcript_id_2.index(remove_marker[l])
         print(index)
         del transcript_id_2[index]
         del lines_tmp2[index]
         count += 1


#write into file
of = open(output_filename, "w")
for i in range(len(lines_tmp2)):
    of.write(lines_tmp2[i])
of.close()
print("\nOutput file: " + output_filename)
```

Supplementary file 3. R script for the heatmap and clustering analysis

```
library (gplots)

data <- read.delim ("Nina_for_heatmap.txt")

rnames <- data[,1]

mat_data <- data.matrix(data[,2:ncol(data)])

rownames(mat_data) <- rnames

hr <- hclust(as.dist(1-cor(t(mat_data), method="pearson")),
method="complete")

colorRampPalette(c("blue","yellow")) -> colour.gradient

heatmap.2(mat_data, col=colour.gradient, breaks=seq(from=-8,to=8,
by=0.001), Rowv=as.dendrogram(hr), Colv=FALSE,
          scale="none", dendrogram="none", key=T, keysize=1,
density.info="none", hclust=function(x) hclust(x,method="complete"),
          distfun=function(x) as.dist((1-cor(t(x)))/2),
          trace="none",cexCol=1.2, labRow=NA)

data$clusternumber <- cutree (hr, 20)

write.table(data, "Nina_clusters.txt")
```

Supplementary file 4. Python script used to extract sequences (written by Nikolas Tolar)

```
### NON GTF

### Nikolas Tolar data extraction tool, at JCU 2019

# ----- Editable part -----


genes_name = 'Mus.fa'
annotation_name = 'Nina_wide_promoters_all.txt'
output = open('Nina_wide_cluster_16.txt','a')
query = open('cluster_16.txt')
merge = 0


'''
    HINT: always edit strings in between the '' symbols

    genes_name = files containing raw DNA sequence - file names should
follow the
                 pattern Xiiii where X is number/letter of chromosome and
                 iiii is the actual name that is shared with all other
files.

                 Variable genes_name holds the part iiii that is shared

    annotation = file containing names of probes and corresponding
locations etc.

    output_file = name of the file the results will save into (if existing
then results will append, otherwise new file will be created)

    transcript_name = name of target transcript

    output_header = header of output file (FASTA format)

    merge = 1 means that the probes will be merged (connected) together
            0 means that the probes will be separated

'''

# ----- Do-not-touch-me part -----

def caller(value,neg,k=0):
    ret = ''

    if neg == 0:
        ret = ret + '_positive_strand_oc_' + str(k) + '\n'
    else:
        ret = ret + '_negative_strand_oc_' + str(k) + '\n'

    return ret

def translate_read_back(string):

    string_new = string[len(string)-1:0:-1] + string[0]

    string_new = string_new.replace('A','R')
    string_new = string_new.replace('T','A')
    string_new = string_new.replace('R','T')

    string_new = string_new.replace('C','F')
    string_new = string_new.replace('G','C')
    string_new = string_new.replace('F','G')

    return string_new


def data_extraction(text, gene_pool):

    start = int(text[2])
    stop = int(text[3])

    segment = gene_pool[start-1:stop]

    return segment

def insert_newlines(string, every=60):
    lines = []

    for i in range(0, len(string), every):
        lines.append(string[i:i+every])

    ret = '\n'.join(lines)
    return ret


def get_exons(genes_name, annotation_name, query, merge):

    transcript_name = query.readline().strip('\n')
    while transcript_name != '':

        annotation = open(annotation_name)
        neg = 0
        res_exons = ''
        res_list = []

        while True:


            text = annotation.readline()
            if text == '':
                break
            if transcript_name in text:
                text = text.split()
# accesing correct chromosome file
                genes = open(text[1]+genes_name)
                genes.readline()
                gene_pool = genes.read()
                gene_pool = ''.join(gene_pool.split())
                genes.close()

                if text[4] == '-':
                    neg = 1

                if merge == 1:
                    res_exons = res_exons +
data_extraction(text,gene_pool)
```

```
                elif merge == 0:

                    res_list.append(data_extraction(text,gene_pool))

        if merge == 1:

            if neg == 1:
                res_exons = translate_read_back(res_exons)

            res_exons = insert_newlines(res_exons)

            message = caller(merge,neg)

            print('>_' + transcript_name + message + res_exons + '\n')
            output.write('>_' + transcript_name + message + res_exons +
'\n\n')

        else:

            for n in range(len(res_list)):
                message = caller(merge,neg,n)

                if neg == 1:
                    res = '>_' + transcript_name + message +
insert_newlines(translate_read_back(res_list[n]))

                else:
                    res = '>_' + transcript_name + message +
insert_newlines(res_list[n])

                print(res + '\n')
                output.write(res + '\n\n')


        annotation.close()
        transcript_name = query.readline().strip('\n')




get_exons(genes_name, annotation_name, query, merge)



output.close()
query.close()
```

Supplementary table 1. List of datasets used in this project

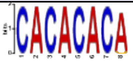
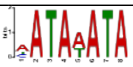
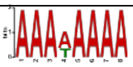
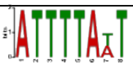
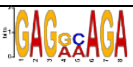
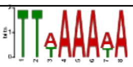
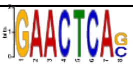
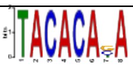
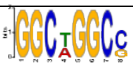
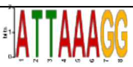
Publication	Cell type	Accession code	RNA type	Mouse strain	Full link
Veselovska et al. (2015)	d5 oocytes	GSE70116	total RNA	C57BL/6Babr	https://www.ncbi.nlm.nih.gov/pubmed/26408185
	d10 oocytes		total RNA	C57BL/6Babr	
	d15 oocytes		total RNA	C57BL/6Babr	
	GV oocytes		total RNA	C57BL/6Babr	
Zhang et al. (2016)	d10 oocytes	GSE71434	polyA RNA	C57BL/6N	https://www.ncbi.nlm.nih.gov/pubmed/27626382
	d14 oocytes		polyA RNA	C57BL/6N	
	GV oocytes		polyA RNA	C57BL/6N	
	MII oocytes		polyA RNA	C57BL/6N	
	zygote		polyA RNA	C57BL/6N x PWK	
	early 2C embryo		polyA RNA	C57BL/6N x PWK	
	late 2C embryo		polyA RNA	C57BL/6N x PWK	
	4C embryo		polyA RNA	C57BL/6N x PWK	
	8C embryo		polyA RNA	C57BL/6N x PWK	
	32C embryo - ICM		polyA RNA	C57BL/6N x PWK	
Wang et al. (2018)	MII oocytes	GSE98150	total RNA	B6D2F1 (C57BL/6 x DB/2)	
	2C embryo		total RNA	B6D2F1 (C57BL/6 x DB/2)	
	4C embryo		total RNA	B6D2F1 (C57BL/6 x DB/2)	
	8C embryo		total RNA	B6D2F1 (C57BL/6 x DB/2)	
	morula embryo		total RNA	B6D2F1 (C57BL/6 x DB/2)	
	E3.5 - ICM		total RNA	B6D2F1 (C57BL/6 x DB/2)	
	E3.5 - TE		total RNA	B6D2F1 (C57BL/6 x DB/2)	
	E6.5 - Epi		total RNA	B6D2F1 (C57BL/6 x DB/2)	
	E6.5 - Exe		total RNA	B6D2F1 (C57BL/6 x DB/2)	
Zhang et al. (2018)	E3.5 - ICM	GSE76505	polyA RNA	C57BL/6N x DBA/2N	https://www.ncbi.nlm.nih.gov/pubmed/29203909 https://www.ncbi.nlm.nih.gov/pubmed/28806168
	E3.5 - TE		polyA RNA	C57BL/6N x DBA/2N	
	E4.0 - ICM		polyA RNA	C57BL/6N x DBA/2N	
	E5.5 - Epi		polyA RNA	C57BL/6N x DBA/2N	
	E5.5 - VE		polyA RNA	C57BL/6N x DBA/2N	
	E6.5 - Epi		polyA RNA	C57BL/6N x DBA/2N	
	E6.5 - VE		polyA RNA	C57BL/6N x DBA/2N	
	Ectoderm		polyA RNA	C57BL/6N x DBA/2N	
	Mesoderm		polyA RNA	C57BL/6N x DBA/2N	
	Endoderm		polyA RNA	C57BL/6N x DBA/2N	
	Primitive_streak		polyA RNA	C57BL/6N x DBA/2N	

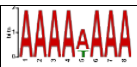
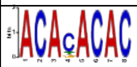
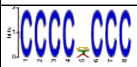
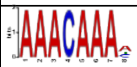
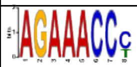
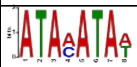
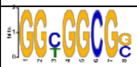
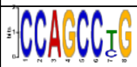
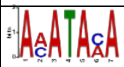
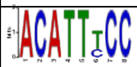
Andergassen et al. (2017)	ESCs	GSE75957	total RNA	FVB/NJxCAST/EiJ	
	E12.5_liver		total RNA	FVB/NJxCAST/EiJ	
	E16.5_liver		total RNA	FVB/NJxCAST/EiJ	
	E16.5_brain		total RNA	FVB/NJxCAST/EiJ	
	E16.5_heart		total RNA	FVB/NJxCAST/EiJ	
	E9.5_VE		total RNA	FVB/NJxCAST/EiJ	
	E12.5_VE		total RNA	FVB/NJxCAST/EiJ	
	E16.5_VE		total RNA	FVB/NJxCAST/EiJ	
	E12.5_placenta		total RNA	FVB/NJxCAST/EiJ	
	E16.5_placenta		total RNA	FVB/NJxCAST/EiJ	
	D3_tongue		total RNA	FVB/NJxCAST/EiJ	
	D3_brain		total RNA	FVB/NJxCAST/EiJ	
	adult_brain		total RNA	FVB/NJxCAST/EiJ	
	adult_liver		total RNA	FVB/NJxCAST/EiJ	
	adult_heart		total RNA	FVB/NJxCAST/EiJ	
	adult_lung		total RNA	FVB/NJxCAST/EiJ	
	adult_spleen		total RNA	FVB/NJxCAST/EiJ	
	adult_thymus		total RNA	FVB/NJxCAST/EiJ	
	adult_leg_muscle		total RNA	FVB/NJxCAST/EiJ	
	adult_virgin_mammary_gland		total RNA	FVB/NJxCAST/EiJ	
	lactating_mammary_gland		total RNA	FVB/NJxCAST/EiJ	
	lactating_brain		total RNA	FVB/NJxCAST/EiJ	
Hanna et al. (2019)	E7.5_Epi	GSE124216	polyA RNA	C57BL/6Babr x CAST	https://www.ncbi.nlm.nih.gov/pubmed/31665063/
	E7.5_Exe		polyA RNA	C57BL/6Babr x CAST	

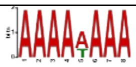
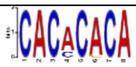
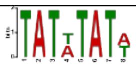
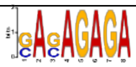
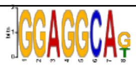
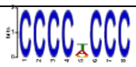
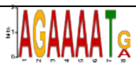
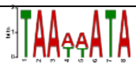
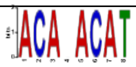
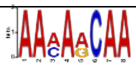
Supplementary table 3. List of imprinted regions in the format chromosome:start-end

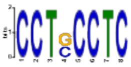
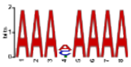
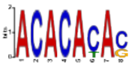
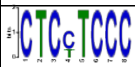
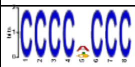
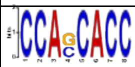
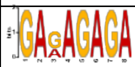
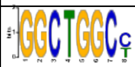
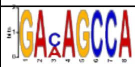
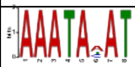
Mouse imprinted regions	
1:63180487-63445890	2:168768109-169633012
10:13009184-13499539	2:174123071-174415803
10:96622810-97565127	3:102206267-102720230
11:11808963-14599275	3:108101433-108148320
11:119040970-119267886	3:41083047-41626719
11:22519235-22990518	4:150652175-150897133
11:51072800-51253650	5:135251231-13535324
11:80968706-81197914	5:18360356-20758662
11:97576186-97627388	5:35615353-35697179
12:109028453-110447119	5:88783282-88886817
13:108407783-110054186	6:30693750-31356742
14:73596143-74732296	6:3603532-5483350
15:100687920-100761746	6:58905233-58907076
15:72034228-73090391	7:102096865-102119397
15:96699699-97244073	7:110639359-110850606
17:3696262-5841327	7:128546980-128696440
17:7011300-14829330	7:142540748-144838082
18:12941841-13006989	7:25754758-25802474
19:38819238-38930914	7:58829421-62778422
19:50778663-52943416	7:6571402-6995299
2:10256530-11172107	8:1198769956-124369048
2:105017905-105224319	8:80739498-80980732
2:122461138-122681232	8:88751946-90247039
2:152635199-152736250	9:107903140-107928468

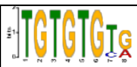
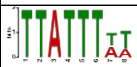
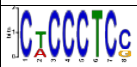
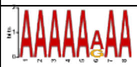
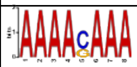
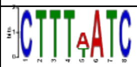
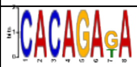
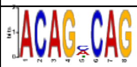
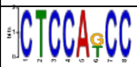
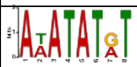
Supplementary table 4. Summary of sequence motifs analysis results

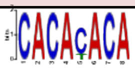
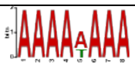
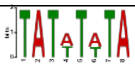
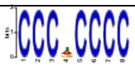
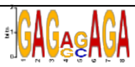
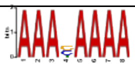
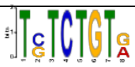
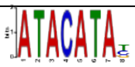
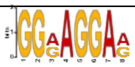
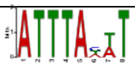
		Sequence	Logo	E-value	Positives	Negatives	Factors
narrow_cl1	motif1	CACACACR		2.20E-18	104 / 180	18 / 180	UP00034_2 (Sox7_secondary) MA1107.1 (KLF9) UP00026_2 (Zscan4_secondary) MA0493.1 (Klf1) GLI2_DBD_1
	motif2	HATAWATA		4.10E-17	120 / 180	32 / 180	FOXC1_DBD_1 UP00094_2 (Zfp128_secondary) CPEB1_full UP00029_1 (Tbp_primary) FOXD2_DBD_1
	motif3	AAAWAAAA		1.40E-14	148 / 180	66 / 180	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00090_2 (Elf3_secondary) UP00028_2 (Tcfap2e_secondary) UP00058_2 (Tcf3_secondary)
	motif4	ATTTTAWT		1.90E-11	76 / 180	12 / 180	UP00121_1 (Hoxd10_2368.2) UP00217_1 (Hoxa10_2318.1) UP00180_1 (Hoxd13_2356.1) UP00078_1 (Arid3a_primary) UP00133_1 (Cdx2_4272.1)
	motif5	GAGRMAGA		1.20E-10	146 / 180	74 / 180	No maches
	motif6	TTWAAAWA		3.70E-09	124 / 180	54 / 180	UP00077_2 (Srf_secondary) MA1125.1 (ZNF384)
	motif7	GAACTCAS		5.60E-09	96 / 180	30 / 180	MA0693.2 (VDR) RARA_full_2 VDR_full UP00064_2 (Sox18_secondary)
	motif8	TACACABA		7.10E-09	97 / 180	31 / 180	UP00034_2 (Sox7_secondary) MA0481.2 (FOXP1) ZSCAN4_full MA1155.1 (ZSCAN4)
	motif9	GGCWGGCS		1.30E-08	100 / 180	34 / 180	ZNF306_full MA1100.1 (ASCL1) Hic1_DBD_1 Hic1_DBD_2 MA0739.1 (Hic1)
	motif10	ATTAAAGG		3.10E-08	70 / 180	14 / 180	MA0151.1 (Arid3a) Tcf7_DBD MA0769.1 (Tcf7) TCF7L1_full MA1421.1 (TCF7L1)

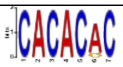
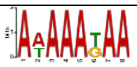
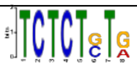
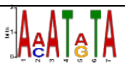
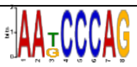
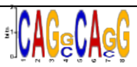
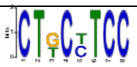
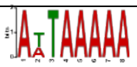
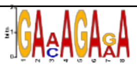
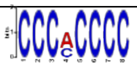
narrow_cl2	motif1	AAAAWAAA		5.10E-19	111 / 130	33 / 130	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00058_2 (Tcf3_secondary) UP00090_2 (Elf3_secondary) UP00028_2 (Tcfap2e_secondary)
	motif2	ACABACAC		4.50E-14	79 / 130	13 / 130	MA1107.1 (KLF9) UP00042_2 (Gm397_secondary) ZSCAN4_full MA1155.1 (ZSCAN4) UP00034_2 (Sox7_secondary)
	motif3	CCCCDCCC		1.30E-08	93 / 130	35 / 130	MA0599.1 (KLF5) UP00099_2 (Ascl2_secondary) MA0079.3 (SP1) SP1_DBD UP00043_2 (Bcl6b_secondary)
	motif4	AAACAAAH		2.00E-07	96 / 130	41 / 130	UP00041_1 (Foxj1_primary) Foxj3_DBD_4 UP00061_2 (Foxl1_secondary) FOXJ2_DBD_3 MA0481.2 (FOXP1)
	motif5	AGAAACCY		2.90E-06	65 / 130	17 / 130	UP00232_1 (Dobox4_3956.2)
	motif6	ATAMATAW		2.90E-06	65 / 130	17 / 130	POU3F3_DBD_3 FOXC2_DBD_2 POU2F3_DBD_2 Foxc1_DBD_2 POU2F1_DBD_2
	motif7	GGYGGCGS		4.50E-06	60 / 130	14 / 130	MA0599.1 (KLF5) MA0079.3 (SP1) MA1102.1 (CTCFL) CTCF_full UP00007_1 (Egr1_primary)
	motif8	CCAGCCYG		1.30E-05	63 / 130	17 / 130	Hic1_DBD_1 UP00035_1 (Hic1_primary) GCM1_full_2 MA0646.1 (GCM1) GLI2_DBD_2
	motif9	AMATAMA		1.40E-05	117 / 130	73 / 130	FOXC2_DBD_2 FOXC1_DBD_1 Foxc1_DBD_1 FOXL1_full_2 FOXJ3_DBD_3
	motif10	ACATTYCC		2.20E-05	46 / 130	7 / 130	UP00013_1 (Gabpa_primary) FLI1_full_1 MA0475.2 (FLI1) Tp53_DBD_3 MA0106.3 (TP53)

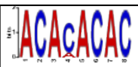
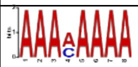
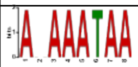
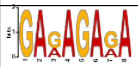
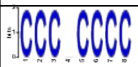
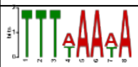
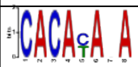
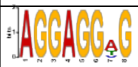
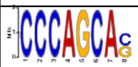
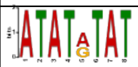
narrow_cl3	motif1	AAAAWAAA		1.40E-41	369 / 447	164 / 447	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00058_2 (Tcf3_secondary) UP00090_2 (Elf3_secondary) UP00028_2 (Tcfap2e_secondary)
	motif2	AAAAWAAA		3.10E-40	301 / 447	96 / 447	UP00034_2 (Sox7_secondary) GLI2_DBD_1 MA1107.1 (KLF9) UP00026_2 (Zscan4_secondary) ZSCAN4_full
	motif3	TATWTATW		1.60E-27	246 / 447	79 / 447	UP00094_2 (Zfp128_secondary) UP00029_1 (Tbp_primary) MEF2B_full MEF2D_DBD MA0660.1 (MEF2B)
	motif4	TATWTATW		1.40E-26	338 / 447	168 / 447	UP00080_2 (Gata5_secondary) MA0482.1 (Gata4)
	motif5	GGAGGCAK		5.70E-22	241 / 447	89 / 447	YY2_full_2
	motif6	CCCCDCCC		1.50E-21	261 / 447	107 / 447	MA0599.1 (KLF5) MA0079.3 (SP1) UP00099_2 (Ascl2_secondary) SP1_DBD UP00043_2 (Bcl6b_secondary)
	motif7	AGAAAATR		6.70E-19	252 / 447	107 / 447	MA0517.1 (STAT1::STAT2)
	motif8	TAAWWATA		7.80E-19	240 / 447	97 / 447	MEF2A_DBD MA0052.3 (MEF2A) MEF2B_full MEF2D_DBD MA0660.1 (MEF2B)
	motif9	ACANACAT		3.90E-18	278 / 447	133 / 447	MA0041.1 (Foxd3) UP00041_1 (Foxj1_primary) Foxc1_DBD_2
	motif10	AAMARCAA		3.50E-17	339 / 447	200 / 447	UP00037_1 (Zfp105_primary) MA0614.1 (Foxj2) UP00025_2 (Foxk1_secondary) FOXJ3_DBD_1 Foxj3_DBD_3

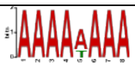
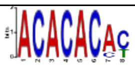
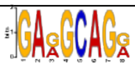
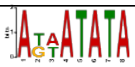
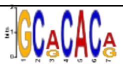
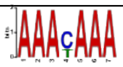
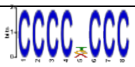
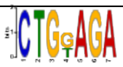
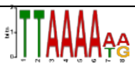
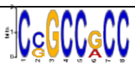
narrow_cl4	motif1	CCTSCCTC		5.30E-10	69 / 101	16 / 101	UP00050_2 (Bhlhb2_secondary) MA0471.1 (E2F6) MA0079.3 (SP1) ZNF784_full
	motif2	AAAHAAAA		2.40E-09	87 / 101	36 / 101	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00090_2 (Elf3_secondary) UP00028_2 (Tcfap2e_secondary) UP00058_2 (Tcf3_secondary)
	motif3	ACACAYAS		1.10E-07	74 / 101	25 / 101	MA1107.1 (KLF9) ZSCAN4_full MA1155.1 (ZSCAN4) UP00042_2 (Gm397_secondary) UP00034_2 (Sox7_secondary)
	motif4	CTCYTCCC		5.80E-07	62 / 101	16 / 101	MA0516.1 (SP2) MA0528.1 (ZNF263) UP00070_2 (Gcm1_secondary) MA0079.3 (SP1) MA0599.1 (KLF5)
	motif5	CCCCDCCC		8.80E-07	75 / 101	28 / 101	MA0599.1 (KLF5) UP00099_2 (Ascl2_secondary) MA0079.3 (SP1) SP1_DBD MA0493.1 (Klf1)
	motif6	CCASCACC		5.90E-05	54 / 101	14 / 101	MA0138.2 (REST) ZBTB7A_DBD ZBTB7B_full MA0694.1 (ZBTB7B) ZBTB7C_full
	motif7	GARAGAGA		2.80E-04	61 / 101	21 / 101	MA0508.2 (PRDM1)
	motif8	GGCTGGCY		3.70E-04	57 / 101	18 / 101	ZNF306_full Hic1_DBD_1 Hic1_DBD_2 MA0739.1 (Hic1) MA0505.1 (Nr5a2)
	motif9	GAMAGCCA		3.20E-04	49 / 101	12 / 101	UP00258_1 (Tgif2_3451.1) YY2_DBD MA0748.1 (YY2) ZNF713_full MA0513.1 (SMAD2::SMAD3::SMAD4)
	motif10	AAATAHAT		7.00E-05	65 / 101	23 / 101	FOXC1_DBD_1 FOXC2_DBD_2 FOXL1_full_2 Foxc1_DBD_1 UP00058_2 (Tcf3_secondary)

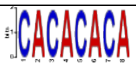
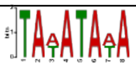
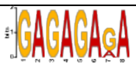
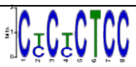
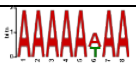
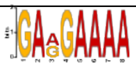
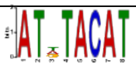
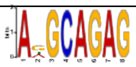
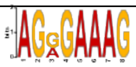
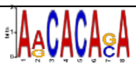
narrow_cl5	motif1	TGTGTGYR		5.90E-13	76 / 101	17 / 101	UP00034_2 (Sox7_secondary) MA1107.1 (KLF9) UP00026_2 (Zscan4_secondary) MA0493.1 (Klf1) ZSCAN4_full
	motif2	TTATTTWW		3.00E-11	75 / 101	19 / 101	MEF2A_DBD MA0052.3 (MEF2A) FOXC2_DBD_2 FOXC1_DBD_1 MEF2B_full
	motif3	CWCCCTCS		8.00E-10	71 / 101	18 / 101	MA0039.3 (KLF4) MA0471.1 (E2F6) MA0057.1 (MZF1(var.2)) MA0528.1 (ZNF263) MA0470.1 (E2F4)
	motif4	AAAAARAA		3.20E-09	91 / 101	42 / 101	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00028_2 (Tcfap2e_secondary) UP00090_2 (Elf3_secondary) UP00058_2 (Tcf3_secondary)
	motif5	AAAASAAA		6.00E-07	68 / 101	21 / 101	MA0442.2 (SOX10) MA0514.1 (Sox3) UP00061_2 (Foxl1_secondary) MA1152.1 (SOX15) UP00039_2 (Foxj3_secondary)
	motif6	CTTTWATC		2.80E-05	44 / 101	7 / 101	UP00029_2 (Tbp_secondary) MA0151.1 (Arid3a) TCF7L1_full MA1421.1 (TCF7L1) Tcf7_DBD
	motif7	CACAGAKA		4.20E-05	53 / 101	13 / 101	MA0140.2 (GATA1::TAL1) FOXB1_DBD_1 UP00080_2 (Gata5_secondary)
	motif8	ACAGHCAG		5.40E-05	63 / 101	21 / 101	MA0513.1 (SMAD2::SMAD3::SMAD4) UP00258_1 (Tgif2_3451.1)
	motif9	CTCCAKCC		1.00E-04	56 / 101	16 / 101	MA1121.1 (TEAD2) MA0471.1 (E2F6) UP00033_1 (Zfp410_primary) MA0470.1 (E2F4) ZNF410_DBD
	motif10	AWATATRT		3.90E-05	47 / 101	9 / 101	UP00094_2 (Zfp128_secondary) NEUROG2_full MA0669.1 (NEUROG2) NEUROG2_DBD UP00029_1 (Tbp_primary)

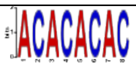
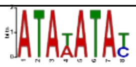
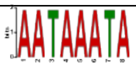
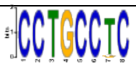
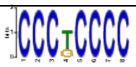
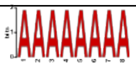
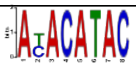
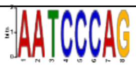
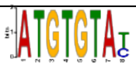
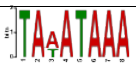
narrow_cl6	No enriched motif found						
narrow_cl7	motif1	CACAYACA		3.90E-35	245 / 406	66 / 406	UP00034_2 (Sox7_secondary) MA1107.1 (KLF9) UP00026_2 (Zscan4_secondary) MA0493.1 (Klf1) ZSCAN4_full
	motif2	AAAAWAAA		1.00E-34	346 / 406	170 / 406	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00058_2 (Tcf3_secondary) UP00090_2 (Elf3_secondary) UP00028_2 (Tcfap2e_secondary)
	motif3	TATWTWTA		1.20E-24	272 / 406	114 / 406	MEF2B_full MEF2D_DBD MA0660.1 (MEF2B) MA0773.1 (MEF2D) UP00094_2 (Zfp128_secondary)
	motif4	CCCDCCCC		4.50E-23	230 / 406	81 / 406	MA0079.3 (SP1) MA0599.1 (KLF5) UP00043_2 (Bcl6b_secondary) SP1_DBD MA0516.1 (SP2)
	motif5	GAGRSAGA		2.70E-22	293 / 406	142 / 406	No matches
	motif6	AAABAAAA		6.90E-18	333 / 406	203 / 406	Foxj3_DBD_4 UP00061_2 (Foxl1_secondary) UP00041_1 (Foxj1_primary) MA1152.1 (SOX15) FOXJ2_DBD_3
	motif7	TSTCTGTR		3.90E-16	268 / 406	136 / 406	MA0002.2 (RUNX1) UP00034_2 (Sox7_secondary) FOXB1_DBD_1
	motif8	ATACATAB		1.20E-15	175 / 406	58 / 406	UP00094_2 (Zfp128_secondary) BHLHE22_DBD MA0817.1 (BHLHE23)
	motif9	GGRAGGAR		3.80E-14	270 / 406	145 / 406	MA0149.1 (EWSR1-FLI1) MA0528.1 (ZNF263) ELF3_full MA0640.1 (ELF3) UP00050_2 (Bhlhb2_secondary)
	motif10	ATTTAHWT		4.10E-12	292 / 406	176 / 406	FOXC1_DBD_1 FOXB1_DBD_3 FOXC1_DBD_3 MA0032.2 (FOXC1) MA0845.1 (FOXB1)

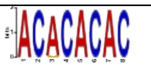
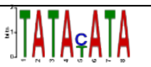
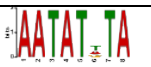
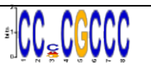
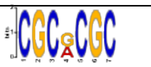
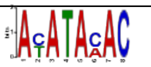
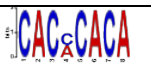
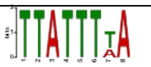
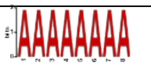
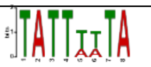
narrow_cl8	motif1	CACACRC		6.50E-20	149 / 195	50 / 195	UP00042_2 (Gm397_secondary) ZSCAN4_full MA1155.1 (ZSCAN4) GLI2_DBD_1 UP00034_2 (Sox7_secondary)
	motif2	AWAAAKAA		7.00E-21	50 / 195	80 / 195	MEF2A_DBD MA0052.3 (MEF2A) UP00073_2 (Foxa2_secondary) MEF2D_DBD MA0773.1 (MEF2D)
	motif3	TCTCTSTR		4.90E-16	163 / 195	75 / 195	FOXB1_DBD_1 MA0140.2 (GATA1::TAL1)
	motif4	AMATRTA		6.30E-15	180 / 195	103 / 195	UP00025_1 (Foxk1_primary) UP00061_1 (Foxl1_primary) ZNF232_full
	motif5	AAKCCCAG		9.90E-14	114 / 195	32 / 195	MA0038.1 (Gfi1) MA0483.1 (Gfi1b) PITX1_full_2 MA0682.1 (Pitx1) PITX3_DBD
	motif6	CAGSCASG		6.40E-13	135 / 195	52 / 195	MA1114.1 (PBX3)
	motif7	CTKCYTCC		8.70E-13	161 / 195	81 / 195	SPDEF_DBD_3 MA0528.1 (ZNF263) ETV6_full_1
	motif8	AWTAAAAA		1.30E-11	125 / 195	46 / 195	CPEB1_full MSX1_DBD_1 HOXA13_full_1 MA0650.1 (HOXA13) Hoxc10_DBD_2
	motif9	GAMAGARA		9.30E-11	153 / 195	77 / 195	MA0508.2 (PRDM1)
	motif10	CCCMCCCC		2.30E-10	99 / 195	28 / 195	ZNF740_full ZNF740_DBD MA0753.1 (ZNF740) UP00021_1 (Zfp281_primary) Zfp740_DBD

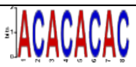
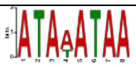
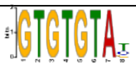
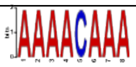
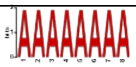
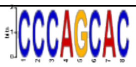
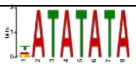
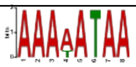
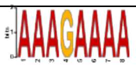
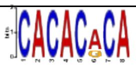
narrow_cl9	motif1	ACAMACAC		1.50E-23	129 / 217	23 / 217	MA1107.1 (KLF9) UP00042_2 (Gm397_secondary) ZSCAN4_full MA1155.1 (ZSCAN4) UP00034_2 (Sox7_secondary)
	motif2	AAAMAAAA		1.60E-18	170 / 217	69 / 217	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00028_2 (Tcfap2e_secondary) UP00090_2 (Elf3_secondary) UP00058_2 (Tcf3_secondary)
	motif3	ANAAATAA		7.10E-14	145 / 217	55 / 217	FOXC2_DBD_2 MEF2A_DBD MA0052.3 (MEF2A) FOXC1_DBD_1 MEF2D_DBD
	motif4	GARAGARA		1.70E-13	170 / 217	82 / 217	MA0508.2 (PRDM1)
	motif5	CCCNCCCC		3.40E-13	171 / 217	84 / 217	MA0599.1 (KLF5) MA0079.3 (SP1) UP00043_2 (Bcl6b_secondary) SP1_DBD MA0516.1 (SP2)
	motif6	TTTWAAWA		3.30E-11	147 / 217	64 / 217	MA1125.1 (ZNF384)
	motif7	CACAYANA		1.60E-09	164 / 217	87 / 217	UP00026_1 (Zscan4_primary) UP00034_2 (Sox7_secondary) UP00026_2 (Zscan4_secondary) MA1107.1 (KLF9) UP00014_2 (Sox17_secondary)
	motif8	AGGAGGHG		3.90E-09	123 / 217	48 / 217	MA0528.1 (ZNF263) UP00057_2 (Zic2_secondary) UP00102_2 (Zic1_secondary)
	motif9	CCCAGCAS		2.40E-08	118 / 217	46 / 217	MA0591.1 (Bach1::Mafk) UP00057_2 (Zic2_secondary) MA0144.2 (STAT3) UP00102_2 (Zic1_secondary)
	motif10	CCCAGCAS		3.40E-08	62 / 217	9 / 217	UP00094_2 (Zfp128_secondary) UP00029_1 (Tbp_primary)

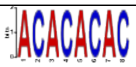
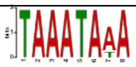
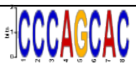
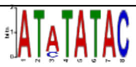
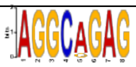
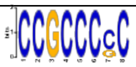
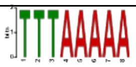
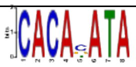
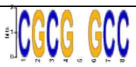
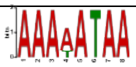
narrow_cl10	motif1	AAAAWAAA		1.10E-16	121 / 147	42 / 147	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00058_2 (Tcf3_secondary) UP00090_2 (Elf3_secondary) UP00028_2 (Tcfap2e_secondary)
	motif2	ACACACMY		1.90E-14	107 / 147	32 / 147	MA1107.1 (KLF9) ZSCAN4_full MA1155.1 (ZSCAN4) UP00034_2 (Sox7_secondary) UP00042_2 (Gm397_secondary)
	motif3	GARGCAGR		7.70E-07	105 / 147	48 / 147	No matches
	motif4	AKWATATA		6.90E-07	57 / 147	10 / 147	No matches
	motif5	GCRCACR		1.70E-06	104 / 147	48 / 147	ZSCAN4_full MA1155.1 (ZSCAN4) UP00042_2 (Gm397_secondary) UP00026_2 (Zscan4_secondary) MTF1_DBD
	motif6	AAAYAAA		1.90E-06	139 / 147	95 / 147	UP00073_1 (Foxa2_primary) MA0851.1 (Foxj3) UP00039_1 (Foxj3_primary) UP00041_1 (Foxj1_primary) UP00025_1 (Foxk1_primary)
	motif7	CCCCDCCC		3.80E-06	104 / 147	49 / 147	MA0599.1 (KLF5) MA0079.3 (SP1) UP00099_2 (Ascl2_secondary) SP1_DBD UP00043_2 (Bcl6b_secondary)
	motif8	CTGKAGA		5.40E-06	117 / 147	64 / 147	No matches
	motif9	TTAAAAWR		1.70E-06	104 / 147	48 / 147	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) MSX1_DBD_1 Msx3_DBD_1 MSX2_DBD_1
	motif10	CSGCCRCC		7.10E-06	59 / 147	13 / 147	UP00007_1 (Egr1_primary) MA0079.3 (SP1) UP00000_2 (Smad3_secondary) UP00002_1 (Sp4_primary) MA0516.1 (SP2)

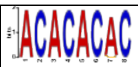
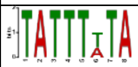
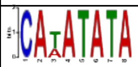
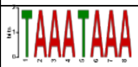
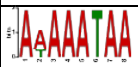
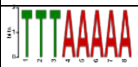
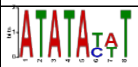
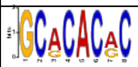
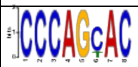
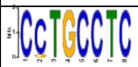
narrow_cl16	motif1	CACACACA		5.60E-49	261 / 540	42 / 540	UP00034_2 (Sox7_secondary) MA1107.1 (KLF9) UP00026_2 (Zscan4_secondary) MA0493.1 (Klf1) ZSCAN4_full
	motif2	TAWATAWA		8.30E-33	413 / 540	208 / 540	UP00094_2 (Zfp128_secondary) UP00029_1 (Tbp_primary) FOXC2_DBD_2 Foxc1_DBD_1 FOXC1_DBD_1
	motif3	GAGAGARA		5.20E-32	325 / 540	123 / 540	UP00011_2 (Irf6_secondary) UP00080_2 (Gata5_secondary)
	motif4	CYCYCTCC		1.10E-26	123 / 540	161 / 540	MA0516.1 (SP2) MA0528.1 (ZNF263) MA0057.1 (MZF1(var.2)) ZNF740_full UP00022_1 (Zfp740_primary)
	motif5	AAAAAWAA		3.70E-26	448 / 540	272 / 540	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00090_2 (Elf3_secondary) UP00028_2 (Tcfap2e_secondary) UP00058_2 (Tcf3_secondary)
	motif6	GARGAAAA		1.20E-27	351 / 540	159 / 540	MA0152.1 (NFATC2)
	motif7	ATDTACAT		3.80E-26	298 / 540	116 / 540	FOXB1_DBD_3 MA0845.1 (FOXB1) FOXB1_DBD_2 FOXC1_DBD_3 MA0032.2 (FOXC1)
	motif8	ADGCAGAG		5.60E-25	324 / 540	142 / 540	No matches
	motif9	AGRGAAAG		3.00E-23	338 / 540	160 / 540	PRDM1_full MA1116.1 (RBPJ) ZNF282_DBD MA1154.1 (ZNF282) UP00086_2 (Irf3_secondary)
	motif10	ARCACASA		2.00E-24	373 / 540	193 / 540	UP00042_2 (Gm397_secondary) UP00026_2 (Zscan4_secondary) UP00025_1 (Foxk1_primary) MA0002.2 (RUNX1) UP00073_1 (Foxa2_primary)

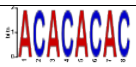
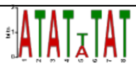
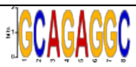
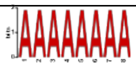
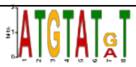
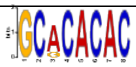
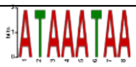
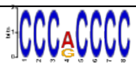
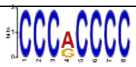
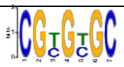
wide_cl1	motif1	ACACACAC		2.50E-33	166 / 180	52 / 180	MA1107.1 (KLF9) UP00042_2 (Gm397_secondary) ZSCAN4_full MA1155.1 (ZSCAN4) UP00034_2 (Sox7_secondary)
	motif2	ATAWATAY		8.90E-26	172 / 180	77 / 180	UP00094_2 (Zfp128_secondary UP00008_2 (Six6_secondary) UP00029_1 (Tbp_primary) FOXD2_DBD_1
	motif3	AATAAATA		3.90E-22	133 / 180	34 / 180	FOXC1_DBD_1 CPEB1_full FOXC2_DBD_2 FOXL1_full_2 Hoxc10_DBD_2
	motif4	CCTGCCKC		4.20E-20	162 / 180	72 / 180	MA0516.1 (SP2) ZNF784_full MA0079.3 (SP1)
	motif5	CCCKCCCC		5.60E-18	158 / 180	71 / 180	MA0079.3 (SP1) UP00033_2 (Zfp410_secondary) MA0599.1 (KLF5) UP00043_2 (Bcl6b_secondary) MA0516.1 (SP2)
	motif6	AAAAAAAA		8.70E-18	178 / 180	110 / 180	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00028_2 (Tcfap2e_secondary) UP00090_2 (Elf3_secondary) UP00058_2 (Tcf3_secondary)
	motif7	AYACATAC		1.20E-17	143 / 180	53 / 180	No matches
	motif8	AATCCCAG		2.70E-15	151 / 180	68 / 180	MA0483.1 (Gfi1b) MA0038.1 (Gfi1) MA0682.1 (Pitx1) PITX1_full_2 PITX3_DBD
	motif9	ATGTGTAY		4.00E-15	132 / 180	47 / 180	MA0613.1 (FOXG1) Foxc1_DBD_2 FOXO1_DBD_2 FOXO1_DBD_1 MA0031.1 (FOXD1)
	motif10	TAWATAAA		9.10E-15	156 / 180	76 / 180	FOXC2_DBD_2 Foxc1_DBD_1 FOXC1_DBD_1 FOXL1_full_2 FOXC2_DBD_3

wide_cl2	motif1	ACRCACAC		1.30E-25	124 / 130	40 / 130	MA1107.1 (KLF9) UP00042_2 (Gm397_secondary) ZSCAN4_full MA1155.1 (ZSCAN4) UP00034_2 (Sox7_secondary)
	motif2	TATAYATA		4.50E-17	97 / 130	22 / 130	UP00094_2 (Zfp128_secondary) UP00029_1 (Tbp_primary) FOXJ3_DBD_3 UP00008_2 (Six6_secondary) FOXB1_DBD_3
	motif3	AATATDTA		5.90E-15	117 / 130	49 / 130	FOXB1_DBD_2 FOXD2_DBD_1 FOXC1_DBD_2 FOXD3_DBD_1 FOXC2_DBD_1
	motif4	CCVCGCCC		6.10E-14	92 / 130	23 / 130	UP00093_1 (Klf7_primary) MA0079.3 (SP1) MA0599.1 (KLF5) SP1_DBD UP00043_2 (Bcl6b_secondary)
	motif5	CGCRCGC		2.20E-13	71 / 130	9 / 130	UP00065_1 (Zfp161_primary) UP00001_1 (E2F2_primary) UP00003_1 (E2F3_primary) MA0632.1 (Tcfl5) MA0506.1 (NRF1)
	motif6	AYATAMAC		6.00E-14	120 / 130	56 / 130	FOXJ3_DBD_3 FOXJ2_DBD_3 Foxj3_DBD_4
	motif7	CACMCACA		6.20E-13	116 / 130	52 / 130	GLI2_DBD_1 MA1107.1 (KLF9) UP00034_2 (Sox7_secondary) UP00026_2 (Zscan4_secondary) ZNF143_DBD
	motif8	TTATTTWA		4.80E-12	122 / 130	64 / 130	ARX_DBD Arx_DBD LMX1B_DBD LMX1A_DBD MA0703.1 (LMX1B)
	motif9	AAAAAAAA		2.30E-11	129 / 130	82 / 130	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00028_2 (Tcfap2e_secondary) UP00090_2 (Elf3_secondary) UP00058_2 (Tcf3_secondary)
	motif10	TATTWWTA		5.30E-11	120 / 130	63 / 130	MEF2A_DBD MA0052.3 (MEF2A) MEF2B_full MA0660.1 (MEF2B) MEF2D_DBD
wide_cl3	No enriched motif found						

wide_cl4	motif1	ACACACAC		3.40E-20	93 / 101	24 / 101	MA1107.1 (KLF9) UP00042_2 (Gm397_secondary) ZSCAN4_full MA1155.1 (ZSCAN4) UP00034_2 (Sox7_secondary)
	motif2	ATAWATAA		3.70E-16	88 / 101	24 / 101	FOXC2_DBD_2 FOXC1_DBD_1 Foxc1_DBD_1 FOXL1_full_2 POU3F3_DBD_2
	motif3	GTGTGTAB		9.30E-14	87 / 101	27 / 101	TBX15_DBD_1 Foxc1_DBD_2 FOXL1_full_1 MA0033.2 (FOXL1)
	motif4	AAAACAAA		3.80E-13	100 / 101	51 / 101	MA0442.2 (SOX10) MA0514.1 (Sox3) UP00061_2 (Foxl1_secondary) MA1152.1 (SOX15) UP00039_2 (Foxj3_secondary)
	motif5	AAAAAAAA		3.10E-12	100 / 101	53 / 101	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00028_2 (Tcfap2e_secondary) UP00090_2 (Elf3_secondary) UP00058_2 (Tcf3_secondary)
	motif6	CCCAGCAC		4.70E-12	86 / 101	29 / 101	MA0591.1 (Bach1::Mafk) UP00096_2 (Sox13_secondary) UP00007_1 (Egr1_primary) ZNF740_full ZNF740_DBD
	motif7	DATATATA		1.90E-11	64 / 101	10 / 101	UP00029_1 (Tbp_primary) UP00094_2 (Zfp128_secondary) UP00008_2 (Six6_secondary)
	motif8	AAAWATAA		7.20E-10	96 / 101	49 / 101	UP00073_2 (Foxa2_secondary) MA1125.1 (ZNF384) MA0497.1 (MEF2C) UP00213_1 (Hoxa9_2622.2) NFATC1_full_1
	motif9	AAAGAAAA		1.80E-09	97 / 101	52 / 101	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00028_2 (Tcfap2e_secondary)
	motif10	CACACRCA		4.30E-09	79 / 101	27 / 101	UP00034_2 (Sox7_secondary) MA1107.1 (KLF9) UP00026_2 (Zscan4_secondary) ZSCAN4_full MA1155.1 (ZSCAN4)

wide_cl5	motif1	ACACACAC		6.30E-19	90 / 101	22 / 101	MA1107.1 (KLF9) UP00042_2 (Gm397_secondary) ZSCAN4_full MA1155.1 (ZSCAN4) UP00034_2 (Sox7_secondary)
	motif2	TAAATAWA		5.20E-14	96 / 101	40 / 101	FOXC2_DBD_2 Foxc1_DBD_1 FOXC1_DBD_1 FOXL1_full_2 UP00073_1 (Foxa2_primary)
	motif3	CCCAGCAC		1.90E-10	85 / 101	31 / 101	MA0591.1 (Bach1::Mafk) UP00096_2 (Sox13_secondary) UP00007_1 (Egr1_primary) ZNF740_full ZNF740_DBD
	motif4	ATMTATAC		2.20E-09	61 / 101	11 / 101	UP00008_2 (Six6_secondary) UP00094_2 (Zfp128_secondary) FOXB1_DBD_3 MA0845.1 (FOXB1) UP00232_1 (Dobox4_3956.2)
	motif5	AGGCRGAG		1.20E-11	97 / 101	47 / 101	MA0065.2 (Pparg::Rxra)
	motif6	CCGCCCSC		9.10E-09	61 / 101	12 / 101	UP00007_1 (Egr1_primary) MA0079.3 (SP1) MA0516.1 (SP2) UP00002_1 (Sp4_primary) KLF16_DBD
	motif7	TTTAAAAA		1.00E-08	94 / 101	48 / 101	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00114_1 (Homez_1063.2)
	motif8	CACABATA		6.60E-08	94 / 101	50 / 101	T_full MA0009.2 (T) UP00026_1 (Zscan4_primary) MA0140.2 (GATA1::TAL1) UP00026_2 (Zscan4_secondary)
	motif9	CGCGNGCC		8.40E-08	47 / 101	5 / 101	UP00001_1 (E2F2_primary) UP00003_1 (E2F3_primary) UP00065_1 (Zfp161_primary) MA1099.1 (Hes1) MA0632.1 (Tcfl5)
	motif10	AAAWATAA		6.60E-09	99 / 101	58 / 101	UP00073_2 (Foxa2_secondary) MA1125.1 (ZNF384) MA0497.1 (MEF2C) UP00213_1 (Hoxa9_2622.2) NFATC1_full_1
wide_cl6	No enriched motif found						
wide_cl7	No enriched motif found						

wide_cl8	motif1	ACACACRC		1.20E-23	117 / 123	38 / 123	UP00042_2 (Gm397_secondary) MA1107.1 (KLF9) ZSCAN4_full MA1155.1 (ZSCAN4) UP00034_2 (Sox7_secondary)
	motif2	TATTTWTA		2.90E-17	118 / 123	52 / 123	MEF2A_DBD MEF2D_DBD MA0052.3 (MEF2A) MA0773.1 (MEF2D) MEF2B_full
	motif3	CAWATATA		1.30E-15	104 / 123	34 / 123	SRF_DBD SRF_full MA0083.3 (SRF) UP00094_2 (Zfp128_secondary)
	motif4	TAAATAAA		1.60E-14	107 / 123	40 / 123	FOXC2_DBD_2 Foxc1_DBD_1 FOXC1_DBD_1 FOXL1_full_2 UP00073_1 (Foxa2_primary)
	motif5	AWAAATAA		4.30E-12	111 / 123	51 / 123	UP00073_2 (Foxa2_secondary) MEF2A_DBD MA0052.3 (MEF2A) MA0497.1 (MEF2C) MA1125.1 (ZNF384)
	motif6	TTTAAAAA		9.30E-12	113 / 123	55 / 123	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00114_1 (Homez_1063.2)
	motif7	ATATAYWT		4.20E-12	109 / 123	48 / 123	UP00094_2 (Zfp128_secondary) UP00029_1 (Tbp_primary) UP00223_2 (Irx3_2226.1) UP00250_1 (Irx5_2385.1) UP00194_1 (Irx4_2242.3)
	motif8	GCRCACRC		1.10E-11	97 / 123	34 / 123	ZSCAN4_full MA1155.1 (ZSCAN4) UP00042_2 (Gm397_secondary) UP00042_1 (Gm397_primary) UP00026_2 (Zscan4_secondary)
	motif9	CCCAGYAC		1.20E-10	108 / 123	50 / 123	UP00096_2 (Sox13_secondary) MA0591.1 (Bach1::Mafk) UP00007_1 (Egr1_primary) ZNF740_full ZNF740_DBD
	motif10	CSTGCCTC		7.00E-11	107 / 123	48 / 123	MA0516.1 (SP2) MA0079.3 (SP1) ZNF784_full
wide_cl9	No enriched motif found						

wide_cl10	motif1	ACACACAC		1.70E-29	133 / 147	34 / 147	MA1107.1 (KLF9) UP00042_2 (Gm397_secondary) ZSCAN4_full MA1155.1 (ZSCAN4) UP00034_2 (Sox7_secondary)
	motif2	ATATWTAT		2.00E-17	112 / 147	31 / 147	UP00094_2 (Zfp128_secondary) UP00029_1 (Tbp_primary) FOXD2_DBD_1 FOXC1_DBD_2 FOXB1_DBD_3
	motif3	GCAGAGGC		2.50E-17	123 / 147	43 / 147	NHLH1_full NHLH1_DBD MA0048.2 (NHLH1) MA0146.2 (Zfx) MA0065.2 (Pparg::Rxra)
	motif4	AAAAAAAA		1.20E-16	145 / 147	82 / 147	MA1125.1 (ZNF384) UP00077_2 (Srf_secondary) UP00028_2 (Tcfap2e_secondary) UP00090_2 (Elf3_secondary) UP00058_2 (Tcf3_secondary)
	motif5	ATGTATRT		4.20E-16	120 / 147	42 / 147	UP00014_2 (Sox17_secondary) UP00051_2 (Sox8_secondary) SOX9_full_3 SOX15_full_3 Sox1_DBD_2
	motif6	GCRCACAC		2.30E-16	113 / 147	34 / 147	UP00042_2 (Gm397_secondary) ZSCAN4_full MA1155.1 (ZSCAN4) UP00042_1 (Gm397_primary) UP00026_2 (Zscan4_secondary)
	motif7	ATAAATAA		3.70E-15	120 / 147	44 / 147	FOXC2_DBD_2 FOXC1_DBD_1 Foxc1_DBD_1 FOXL1_full_2 Foxj3_DBD_3
	motif8	CCCRCCCC		1.40E-14	127 / 147	54 / 147	MA0599.1 (KLF5) UP00043_2 (Bcl6b_secondary) MA0079.3 (SP1) SP1_DBD MA0516.1 (SP2)
	motif9	CCTGTCTC		1.50E-14	118 / 147	43 / 147	MA1114.1 (PBX3) UP00086_2 (Irf3_secondary) MEIS3_DBD_1 MA0775.1 (MEIS3) MA0513.1 (SMAD2::SMAD3::SMAD4)
	motif10	CGYGYGC		2.30E-14	112 / 147	37 / 147	UP00097_1 (Mtf1_primary) MTF1_DBD MA0863.1 (MTF1) MA1099.1 (Hes1) UP00065_1 (Zfp161_primary)
wide_cl16	No enriched motif found						